DOCUMENT RESUME

ED 400 169 SE 058 842

AUTHOR	Jorgensen, Margaret
TITLE	Rethinking Portfolio Assessment: Documenting the Intellectual Work of Learners in Science and Mathematics.
INSTITUTION	ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, Ohio.
SPONS AGENCY	National Science Foundation, Arlington, VA.; Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	96
CONTRACT	MDR-9154422; RR93002013
NOTE	227p.
AVAILABLE FROM	ERIC/CSMEE Publications, The Ohio State University, 1929 Kenny Road, Columbus, OH 43210-1080.
PUB TYPE	Guides - Classroom Use - Teaching Guides (For Teacher) (052) -- Information Analyses - ERIC Clearinghouse Products (071)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	Educational Research; Elementary Education; Junior High Schools; *Mathematics Instruction; Middle Schools; *Portfolio Assessment; Portfolios (Background Materials); *Science Instruction; *Student Evaluation

ABSTRACT

This book details the theory and practice of portfolio assessment in mathematics and science for the elementary and middle grades as implemented in the Authentic Assessment for Multiple Users Project funded by the National Science Foundation. Included in this document are specific assessment tasks, teacher directions for administering these tasks, scoring guides or rubrics for each task, and exemplars of student work for these scoring guides. Chapter 1 provides background information about how the portfolio paradigm associated with the Authentic Assessment for Multiple Users Project compares with other approaches. Chapter 2 chronicles the collaborative journey to consensus of the project participants, and Chapter 3 details the part of the process that yielded assessment strategies. Chapter 4 considers what worked, what worked well, and what didn't work at all. Chapters 5 through 12 present various assessment tasks. Each task includes teacher directions, scoring guides, and support materials with the tasks and their ancillary materials presented in camera-ready form. Chapters 13 and 14 discuss the development of scoring guides and lessons learned from the project process and findings. Contains 20 references. (DDR)

Rethinking Portfolio Assessment

Documenting the Intellectual Work of Learners in Science and Mathematics

Margaret Jorgensen

Rethinking Portfolio Assessment

Documenting the Intellectual Work of Learners in Science and Mathematics

Margaret Jorgensen

ERIC Clearinghouse for Science, Mathematics, and Environmental Education

Columbus, Ohio

1996

Cite as:

Jorgensen, M. (1996). *Rethinking portfolio assessment: Documenting the intellectual work of learners in science and mathematics.* Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Document development:

David L. Haury, *ERIC/CSMEE Executive Editor*
Linda A. Milbourne, *ERIC/CSMEE Copyeditor*
Cover and design by Haury and Milbourne

ERIC Clearinghouse Accession Number: SE 058 842

This document and related publications are available from ERIC/CSMEE Publications, The Ohio State University, 1929 Kenny Road, Columbus, OH 43210-1080. Information on publications and services will be provided upon request.

ERIC/CSMEE invites individuals to submit proposals for monographs and bibliographies relating to issues in science, mathematics, and environmental education. Proposals must include:

- A succinct manuscript proposal of not more than five pages.
- An outline of chapters and major sections.
- A 75-word abstract for use by reviewers for initial screening and rating of proposals.
- A rationale for development of the document, including identification of target audience and the needs served.
- A vita and a writing sample.

This publication was funded in part by the Office of Educational Research and Improvement, U. S. Department of Education under contract no. RR93002013. Opinions expressed in this publication do not necessarily reflect the positions or policies of OERI or the Department of Education.

The project on which this publication is based was funded by the National Science Foundation under Grant No. MDR 9154422. Any opinions, findings, and conclusions or recommendations expressed herein are those of the author and do not necessarily reflect the views of the National Science Foundation or Educational Testing Service.

★★ Contents ★★

ERIC and ERIC/CSMEE

The *Educational Resources Information Center* (ERIC) is a national information system operated by the Office of Educational Research and Improvement in the U. S. Department of Education. ERIC serves the educational community by collecting and disseminating research findings and other information that can be used to improve educational practice. General information about the ERIC system can be obtained from ACCESS ERIC, 1-800-LET-ERIC.

The ERIC *Clearinghouse for Science, Mathematics, and Environmental Education* (ERIC/CSMEE) is one component in the ERIC system and has resided at The Ohio State University since 1966, the year the ERIC system was established. This and the other 15 ERIC clearinghouses process research reports, journal articles, and related documents for announcement in ERIC's index and abstract bulletins.

Reports and other documents not published in journals are announced in *Resources in Education* (RIE), available in many libraries and by subscription from the Superintendent of Documents, U. S. Government Printing Office, Washington, DC 20402. Most documents listed in RIE can be purchased through the ERIC Document Reproduction Service, 1-800-443-ERIC.

Journal articles are announced in *Current Index to Journals in Education* (CIJE). CIJE is also available in many libraries, and can be purchased from Oryx Press, 4041 North Central Avenue, Suite 700, Phoenix, AZ 85012-3399 (1-800-279-ORYX).

The entire ERIC database, including both RIE and CIJE, can be searched electronically online or on CD-ROM.

Online Vendors: BRS Information Technologies, 1-800-289-4277
DIALOG Information Services, 1-800-334-2564
OCLC (Online Computer Library Center), 1-800-848-5800

CD-ROM Vendors: DIALOG Information Services, 1-800-334-2564
Silver Platter Information, Inc., 1-800-343-0064

Researchers, practitioners, and scholars in education are invited to submit relevant documents to the ERIC system for possible inclusion in the database. If the ERIC selection criteria are met, the documents will be added to the database and announced in RIE. To submit, send two legible copies of each document and a completed Reproduction Release Form (available from the ERIC Processing and Reference Facility, 301-258-5500, or any ERIC Clearinghouse) to:

ERIC Processing and Reference Facility
Acquisitions Department
1301 Piccard Dr., Suite 300
Rockville, MD 20850-4305

For documents focusing on science, mathematics, or environmental education, submit two copies of each document, together with a completed *Reproduction Release Form*, directly to:

ERIC/CSMEE Acquisitions
1929 Kenny Road
Columbus, OH 43210-1080

Acknowledgments

The teachers and administrators who volunteered for the Authentic Assessment for Multiple Users project were and are very special people. They jumped into the chasm of uncertainty because they were committed to better forms of assessment. They worked long and hard and taught the project staff a great deal. Working together, they became fast friends and trusted colleagues despite their geographic and situation-specific differences. My sincere appreciation and respect go to each and every one of these individuals.

> Billie Jean Ellington, Steve Ruff, Ellen Marie Moore, Wilma Smith, Tinena Bice, and Darlene Reynolds from Dade County Public Schools. Sherry Gibney, Kim Moore, Sara Glickman, Susan McClendon, and Letitia Blackmon from Clarke County Public Schools. Judy Dennison, Carol J. Cobb, Mamie Anderson, Della Hodges, and Rosilyn Smith from the Fulton County Public Schools. Linda Shearon, Kathy Singleton, Sandra Gouldthorpe, Virginia Gilbert, and Claudia Cook From Marietta City Public Schools. Barbara Reed, Sharon Mahan, Tara Bray, Allen Brumbalow, Debbie Crawford, and Cindy Nesbit from the Gwinnett County Public Schools. Carol Rountree, Carol Fuller, Beth Pearson, April White, and Nancy Boatwright from the Richmond County Public Schools.

And, to help all of us, we had the advantage of outstanding consultants playing a variety of roles. We had Anneli Lax, Richard Lesh, Gerald Wheeler, and Michael Padilla as content advisors. We had Leon and Pearl Paulson as evaluators. We had Nancy Cole, Henry Braun, and Richard Noeth as corporate advisors. Finally, we had a terrific project staff. I was assisted and challenged by Dr. Martha-Anne McDevitt. Dr. Ted Chittenden, Dr. Terry Salinger, Ms. Roberta Camp, and Dr. Roy Hardy added expertise and wisdom. In addition, we were fortunate to have an excellent support team. Without the diligence and demands for high quality of Drucilla Jackson and the careful eye of Katherine Goodman, very little of this good work would have been possible.

All of us owe the National Science Foundation a debt of gratitude for supporting all of us in this wonderful, intellectual, and important project. A special thanks goes to Dr. Francis X. Sutman who served as project monitor during the first half of the project and remains a friend and advisor. Dr. Emma Owens guided the project through to the end with frequent words of support.

Preface

This book details the theory and practice of portfolio assessment in mathematics and science for the elementary and middle grades as implemented in the "Authentic Assessment for Multiple Users" project funded by the National Science Foundation (1992–1994). It presents the new paradigm as an outgrowth of more typical portfolio assessment models including specific assessment tasks, student data, and scoring data in support of this new paradigm. Included in this manuscript are specific assessment tasks, teacher directions for administering these tasks, scoring guides or rubrics for each task, and exemplars of student work for these scoring guides. In toto, these materials may be directly extracted from this manuscript and merged into an assessment system if the reader judges a task to be a valid reflection of what is taught and learned and if the reader judges a task to be a useful, meaningful, and credible tool for collecting evidence about what students think, know, and can do in science and/or mathematics. Issues of performance transfer and generalizability are also addressed. Finally, this book recounts the training strategies used to build consensus among the participants both for the strategies and process of training as well as for the value of the assessment strategies in terms of the transformation of teaching and learning at the classroom level.

MJ

1

Portfolio Assessment in Mathematics and Science

This chapter presents an overview of the project purpose within the context of portfolio assessment as it has been implemented in a variety of situations. The portfolio paradigm associated with the Authentic Assessment for Multiple Users study is contrasted with the approaches commonly reported in the professional literature.

There is no doubt that the tests students take influence what both teachers and students value. "Over the past decade, it has been repeatedly demonstrated that assessment influences what students learn and what teachers teach."[1] It has also been concluded that:

> Classroom evaluation...guides [student] judgment of what is important to learn, affects their motivation and self-perception of competence, structures their approaches to and timing of personal study, consolidates learning, and affects the development of enduring learning strategies and skills. It appears to be one of the most potent forces influencing education.[2]

The project described in these pages resulted from a recognition of the important influence that tests have, not only on what is taught, but what is perceived as important to be taught by teachers and important to be learned by students. It follows naturally that assessments should support important teaching and important learning. Portfolio assessment and performance assessment have these two goals. Advocates of these new forms of assessment argue persuasively that new types of tests can model ideal instruction through their substance and format and that they can capture important evidence about what students think, know, and can do. There is also the implicit expectation that this evidence can be judged against well-articulated standards of quality.

In the literature, portfolio assessments are typically described as:

> Purposeful collections of student work that exhibits the student's efforts, progress, and achievements in one or more areas. The collection must include student participation in selecting contents, the criteria for selection, the criteria for judging merit, and evidence of student self-reflection.[3]

The idiosyncratic nature of portfolio assessment espoused by the Paulsons and others presents the unique and individualistic content of each student's portfolio as a primary interest. However, this perspective makes answering the question, "What is quality work?" literally impossible to answer. In fact, without some common attributes of either stimulus and evidence across learners there is virtually no way to compare the performance of different students one with the other. Likewise, there can be no assessment without some common attributes of either stimulus or evidence between a student or a group of students and the appropriate standards of quality designed to evaluate the work of learners. An approach to meet these pressing demands was crafted in this project by blending a range of structured to ill-structured portfolio entries. Their nature and scope are discussed specifically in Chapters 5–12.

The appeal of portfolio assessment and performance assessment as effective ways to capture meaningful evidence of what students think, know, and can do appeared to be seriously negated by the limited utility of assessments that cannot provide credible information relative to external standards, for instance, normative indicators or criterion-referenced indicators. Thus, the challenge became straightforward: How can the appeal and value of idiosyncratic and individually-sensitive assessment strategies that model ideal instruction and that reflect meaningful learning be maintained within the context of assessments that allow for the elicited evidence to be judged fairly, equitably, and consistently against some stable frame of reference.

---

[1] Moss, P.A., Beck, J. S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1992, Fall). Portfolios, accountability, and an interpretive approach to validity, *Educational Measurement: Issues and Practice, 11*(3), 12-21.

[2] Crooks,T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 85*(4), p. 467.

[3] Paulson, F.L., Paulson, P.R., & Meyer, C.A. (1991, February). What makes a portfolio a portfolio? *Educational Leadership*, pp. 60-63.

The final challenge stemmed from a practical reality that any assessment that takes time, energy, resources, and money will most likely only survive if the assessments inform decisions beyond those made for individual students. In other words, if portfolio assessments (including performance assessments) are only useful in describing individual students, they will not likely be valued by anyone outside that classroom. In short, some level of aggregation must be possible in order for the portfolio assessments to sustain support. Thus, the notion of aggregation (over time or over group) became important.

The portfolio assessment project and its products being presented here have roots in frustrations with the notion of portfolio assessment as presented in the professional literature of 1988 through 1992. Specifically, the notion that assessment can be fair, credible, reliable, and valid while being defined by individual students or students in consultation with their teachers seemed particularly contradictory. On the other hand, the value of collections of student work as evidence of what students had learned, thought about while they learned, and what students could actually do seemed incredibly important. So, too, is the notion of being able to combine (aggregate) information (data) over students and over time. So, in an effort to blend the characteristics of sound measurement practices with the delicious thought of student-centered assessments, the Authentic Assessment for Multiple Users (AAMU) project took shape.

The particular perspective that shaped the AAMU project and this presentation of it reflects the preferences of the author to respect a fundamental property of measurement well-said by Cronbach:

> A test is a systematic procedure for observing behavior and describing it with a numerical or categorical score.[4]

This unique perspective caught the attention of the National Science Foundation (NSF) in 1992 and that organization funded AAMU for three years.

More common were portfolio projects where categories of activities were requested as portfolio entries with the decisions of specific tasks or types of evidence residing with teachers, students, or teachers and students jointly. It was our intention to moderate this approach by increasing the systematic nature of the tasks included in portfolios, thereby increasing the value of the entries as assessments. This focus included offering a range of structured tasks that would accommodate content-specific tasks and tasks for which the content could vary.

Driving the project staff was a belief that Cronbach's definition could provide a platform for tasks that would model ideal instruction, elicit credible and meaningful evidence of learning consistent with the habits of the disciplines, and that would yield sound information about individual learners. Early in this project, the Rand report[5] on the Vermont portfolio study gave credence to our position albeit from a different perspective.

Our question became how to adhere to the fundamental property of systematic assessment and systematic scoring while supporting the important role of the learner in documenting what he or she thinks, knows, and can do. In essence, we pondered the possibility of bridging constructivist learning theory with new ways to systematically document the intellectual work of learners. The Authentic Assessment for Multiple Users project was designed *to determine whether portfolio assessment can be structured to permit meaningful aggregation for multiple hierarchical users.* The content platform was science and mathematics instruction at grades three through six.

We began this project with a liberal definition of what portfolio assessment should be: *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. We then supported this liberal and somewhat ill-structured definition with one often reported in the literature:

---

[4]Cronbach, L. J. (1970). *Essentials of psychological testing.* New York: Harper & Row, p. 26.

[5]Koretz, D., McCaffey, D., Klein, S., Bell, R., & Stecher, B. (1992, December). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program.* Washington, DC: RAND Institute on Education and Training.

> 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[6]

Even though this definition was typically used by proponents of portfolio assessment with roots in the whole language movement, our position is that this definition in and of itself does not preclude systematic ways of collecting and scoring evidence of student learning. Thus, it did not seem contradictory. The challenge was to develop a model for portfolio assessment that would lead to development of purposeful and meaningful work that would be valued by the student, the teacher, and others who had information needs about a student or group of students. It was our view at the onset of this project that if we could build into the portfolio assessment paradigm the characteristics Cronbach uses to define a traditional test, we would surely have the best of both worlds, credible and meaningful assessment information.

What the portfolio would look like was open for creative invention. These collections of purposeful and meaningful evidence of learning could be virtually unlimited in their structure, shape and nature. Limits would stem from the academic discipline itself and from what would be meaningful work in the discipline rather than from preconceived notions of what systematic assessment should look like. Implicit in this notion, however, is that collection, selection, and reflection are still desirable descriptors of the portfolio process.

Theoretical Framework

The theoretical framework for this study (see Figure 1.1) is derived from the work of Paulson and Paulson.[7] Based on the Paulson's early work on portfolios, an extension of an assessment model loosely based on Stake's evaluation model[8] was proposed. Beginning with their Activity, Historical, and Stakeholder dimensions, the principal investigator for the AAMU project reconceptualized the 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 , 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

[6]Arter, J., & Spandel, V. (1991, Spring). NCME Instructional Model: Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, 2(1), pp.36-44.

[7]Paulson, F.L. & Paulson, P.R. (1990, August). *How do portfolios measure up? A cognitive model for assessment portfolios.* Paper presented at a conference, Aggregating Portfolio Data, sponsored by the Northwest Evaluation Association, Union, WA [ERIC Document Reproduction Service ED 324 329].

[8]Stake, Robert E. (1967). *The countenance of educational evaluation*, Springfield, IL: Gifted Children Section, Department of Exceptional Children.

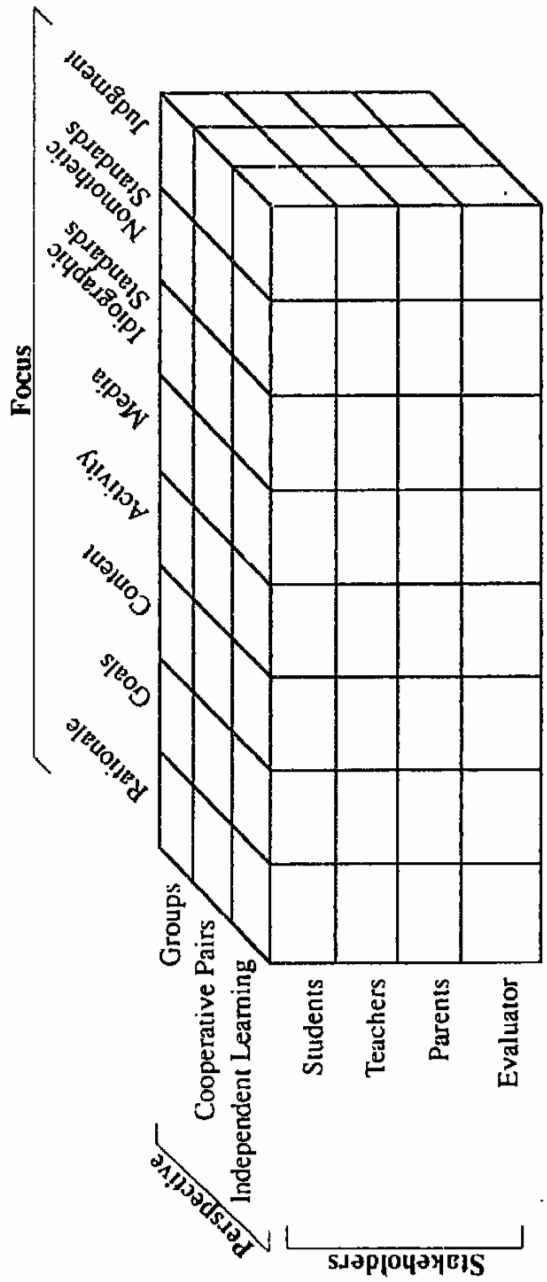


Figure 1.1: AAMU Evaluation Model

consensus about all dimensions of a portfolio design that can be adopted by multiple users in both hierarchical and horizontal environments.

The model 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. One of the goals of this study was to examine the practical utility of this model. Modifications made to the Paulson's model include redefinition of their dimensions as follows:

> 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 provided teachers and developers with a framework for portfolio assessment that was particularly sensitive to the various levels of information users of assessment information. It provided a structure for planning that theoretically optimized the possibility that the assessment would work effectively for multiple users and that its application would produce meaningful aggregate data. Further, this model defined 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 was used in this project to structure a process of consensus building among teachers, students, parents, and evaluators. Each portfolio assessment entry is 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 decision making. Consensus is built regarding the dimensions of the portfolio that are likely to impact meaningful aggregation. For example, the participants in this project were guided through the model with the understanding that the product of their work must be an assessment activity that supports use by each member of the team. This meant that the decisions about what constitutes a portfolio and its purpose(s), when entries were to be made, who would select entries, how they would be "scored," what standards would be used, and how the aggregated portfolio information at the student, classroom, and school levels would be communicated and used must be made by a consensus of users at each level of the model.

In these pages, both the process and products of the AAMU study are reported. The learning journey for all participants was an amazing one, as documented in quotations from the teachers and participating administrators. The products of this study are also notable. They represent the

diversity and richness of evidence one would hope to obtain through an assessment process that empowers the learner to take responsibility for learning. Some assessment tasks are more obviously connected to the habits of mind of the discipline than are others. Some appear to have potential for use across a variety of content areas. Some are more engaging than others. As the story of the learning journey unfolds and as the specific assessment tasks are shared, we hope that the notion of portfolios as true assessment strategies will become feasible and appealing to the reader.

2

Project Partners and the Beginning of Consensus

The notion of teachers as researchers is critical to educational reform in general and certainly to the transformation of assessment in the context of constructivist learning theory. The motivations for participation in this project are, perhaps, more revealing about the conditions that support the risk-taking behavior required for paradigm shifts than of the personality variables of the participating individuals. However, it goes without saying that the vast majority of the participants in this project were reform-minded courageous educators. The thirty-month learning journey was not completed by all beginning participants. For some, the work was too difficult. For others it was too frustrating. For others it was not useful. But, for those who struggled and completed the journey, the impact of their work on their lives as educators is remarkable. This chapter chronicles this learning journey.

From the perspective of measurement as a field of study and as a vocation, one of the most powerful aspects of work in portfolio assessment has been the widespread belief in portfolios held by teachers. The "face validity" of portfolios is typically without doubt. Teachers who have read about portfolios or who have tried them in some degree tend to be extremely supportive of their use and optimistic about their value. There is no question about whether or not meaningful work can be captured through portfolio assessment. Issues of how to score and report the evidence remains, however, of considerable concern.

Project Partners

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

ETS Staff

The project staff included Roberta Camp, Ted Chittenden, Marty McDevitt, Terry Salinger, Margie Jorgensen, Drucilla Jackson, Katherine Goodman, and David Powell. Ms. Camp and Dr. Salinger are 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 with school systems in the area of language arts portfolio assessment. Dr. Chittenden is a science educator, test developer, and consultant in the general area of documentation of student learning for the purpose of informing instruction. Dr. McDevitt and Dr. Jorgensen (principal investigator) are experienced developers in both traditional and innovative types of assessments. Ms. Jackson, Mrs. Goodman, and Mr. Powell provided powerful support for the process and products of this project.

Internal Advisors

The internal advisors included Henry Braun, Vice President for Research Management at ETS, Nancy Cole, currently President of ETS but then Executive Vice President for ETS, and Rick Noeth, Vice President for the Field Service Division. Each of these individuals was involved in the decision to propose this project to NSF and their support for this project was evident in their continuing roles during the life of this project.

External Advisors

The external advisors brought to the project provided unique and important perspectives from outside the measurement community. Dr. Anneli Lax, recently retired from the Courant Institute of Mathematical Sciences at New York University, Dr. Richard Lesh, then a senior research scientist at ETS in the area of mathematics education, and Dr. Michael Padilla, then Chair of the Science Education Department of the University of Georgia were actively involved in guiding this project to insure high quality content as well as assessment design. In addition, upon his retirement from NSF, Dr. Frank X. Sutman (the project's original monitor) joined the project as an advisor. His expertise in particular, both with new forms of assessment and with science content, served the project well.

External Evaluators

Drs. Pearl and Leon Paulson served as external evaluators. As developers of the model upon which this project was based, they provided unique and powerful criticism and insights that challenged and stimulated all project partners. What was lacking in objectivity was more than

offset by their knowledge about portfolios, measurement, and the notion of aggregation as an important outcome of portfolio use.

School Partners

In recruiting school systems that would be interested in participating in this project if funded, it was not surprising that we quickly obtained the number we wanted to work with. Each of the six Georgia school systems that volunteered schools and teachers to participate in this project held the same firm conviction in portfolio assessment described above. Each system was willing and eager to participate, hoping that representatives of the district would learn how to implement a portfolio assessment system that would have the credibility, utility, and value attributed to traditional assessment strategies (i.e., norm-referenced multiple-choice tests).

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 2.1, there is considerable variability in the demographics and financial commitment to education across these systems.

Table 2.1 School Demographics

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 the 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.

The project partners were supported in their work by staff members from Educational Testing Service (ETS). As the project continued over a three-year period, subject area specialists became

more and more important. We found that the teachers and administrators participating in this project were, as typical of elementary teachers, ill-prepared in the content of science and in the content of mathematics. Thus, as they struggled with crafting strategies to capture evidence of the intellectual work of students in science and mathematics, the quality of their ideas was subject to the academic preparation that they had in the discipline. The work of subject area specialists became an important mechanism to enhance the quality of the assessment strategies. We found it necessary to substantially increase contact time between the school teams and subject matter specialists. As a result, science and mathematics experts joined the project as consultants to work directly with the school teams.

We also benefited from a group of external advisers. These individuals represented the disciplines of science or mathematics and lent an objective eye to the product development phase of this project. Their unique and important perspectives brought a degree of content rigor that might otherwise not have been present.

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 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 a research project 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.

The planning meeting was critical in reaffirming each system's commitment to the project. By so doing, each system publicly acknowledged that the teachers and the students who would be participating in the project would 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 seemed trivial, they contributed to the visibility of this project in the local school setting. This visibility was 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 served 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 were building level administrators. One was an instructional lead teacher. The Team Liaison was the primary contact between the ETS project staff and the school teams. The relationship among partners on this project is depicted in Figure 2.1.

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 2.2, the teachers were identified primarily on the basis of their willingness to participate, their instructional expertise, and their commitment to quality and to change.

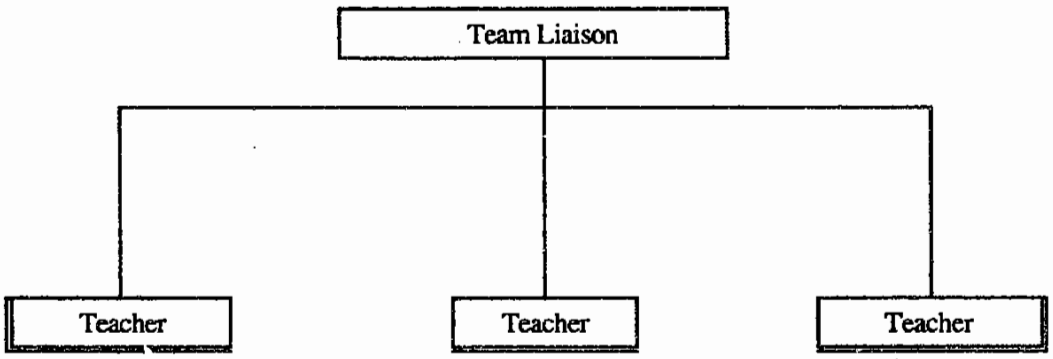


Figure 2.1 School System Team Structure

Table 2.2 Teacher Participants

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	• 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

It is important to recognize that the teachers who chose to become partners with ETS on this project had demonstrated time after time a willingness to take chances, to be creative, and to work very, very hard. And, although five individuals resigned over the course of the project, only two because of disagreements within the project or dissatisfaction with the project or products. However, it is also important to note that these teachers were not exceptionally well-trained in either science or mathematics. Nor did they live or work in communities that supported education

reform, site-based management, or outcomes-based education. They did work for school systems that were open-minded but skeptical about whether or not this approach to student assessment would win favor in the community. It is also evident that during the three years of this project the local mood towards reform and related issues became less tolerant. In fact, one of the teams literally had to stop using the phrase "portfolio assessment" and assume a low-profile as they continued their important work on the project.

The grade-level focus for this project was grades three through six. The content-area focus was science and mathematics or an interdisciplinary or thematic approach to these areas. The emphasis on these areas resulted in the availability of content standards in mathematics and the logical relationship between science and mathematics content and process.

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, 1992. This delay impacted the project rather significantly because of the schedules of the participating school systems.

By the end of the project, the participants were 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 amounted to more than 3500 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 revealed that the teams continued discussion during lunches and other breaks, and on their written *Daily Reflections*.

3

Developing Assessment Strategies and Sustaining Consensus

It is no easy or straightforward process to develop assessment tasks that evoke from learners cognitively complex and meaningful behaviors in support of the National Council of Teachers of Mathematics *Curriculum and Evaluation Standards for School Mathematics* (1989), *Science for All Americans* (1989,1990), and *Benchmarks for Science Literacy* (1993). This chapter describes the process of assessment task creation within the context of this particular project. It also describes the extension of a traditional model of development, revision, and refinement that has been used in traditional test development for decades. The effective use of an iterative process of development, revision, and refinement for this new assessment paradigm provides both a strategic process link to tried and true test development strategies and a systematic framework for forging new ground in assessment without abandoning a critical reflective and analytical model. Here a mapping process is described that causes the developers to re-examine the task and scoring guide in systematic ways at numerous points in its development.

Training Highlights

Participants met for the first time in August 1992 at 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 the Authentic Assessment for Multiple Users project. 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, for example, 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 always included three elements: pre-session reading and journal writing, active participation in large-group, small-group, and individual activities, and post session follow-up assignments for small groups. The training was designed to model ideal instruction and to adhere to constructivist learning theory. Active learning was supported and the participants were

encouraged to become comfortable as an expert resource in some aspect of portfolio assessment so that they could serve as local resident resources back at their school.

For the initial training session, each school-based team was sent a list of guiding questions in advance of the training sessions:

- Based upon the readings and your own experience, craft a definition of portfolio assessment. Include descriptors of what it is and what it is not both in terms of common practice and possibilities.

- In *Science for All Americans*, scientific literacy is presented as a central goal of education. What does this suggest in terms of what school-based learning is and how it occurs? How might portfolio assessment foster this new direction?

- As an assessment tool, what advantages and disadvantages does portfolio assessment offer over other, more traditional assessment tools?

The reading materials sent in advance of the initial training session were selected to stimulate thinking on these and other questions. The reading materials were selected because they represented state-of-the-art assessment approaches in science 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.

During the initial three-day training session, 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 d er/enforcer - students as sponge" paradigm (see Tables 3.1 and 3.2).

The difference in the approaches taken by the two groups suggests two very different philosophies. For Group 1, the emphasis was on the development of learners for a changing and increasingly technological and scientific world. For Group 2, the emphasis was first on establishing the assessment strategy and secondly on empowering students. This difference in perspective led to considerable discussion and rethinking extending to two additional days.

This disagreement in focus highlights one commonly voiced by teachers involved in trying out new forms of assessment; often the assessment format becomes more important than the evidence of student learning. Here also that tendency was expressed; a large portion of the group felt so compelled to make portfolio assessment work that they positioned that as the priority rather than focus on student learning and letting the assessment formats emerge as natural extensions of the teaching/learning process. It was also evident that where representatives from six different school systems might disagree on assessment formats readily, they would be more likely to reach consensus on broad academic expectations. The project staff worked hard to propose a compromise *Rationale* and set of *Goals* which would be adopted by consensus. These were presented to the research partners as reported in Table 3.3:

Table 3.1 Rationale and Goals Identified by 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 3.2 Rationale and Goals Identified by 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.

Table 3.3 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 solutions, and evaluate outcomes (G-1: 1, G-2: 2,4)[1]

- **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 (G-1:3, G-2:1)

- **Effective communicators** (G-2:5)

- **Effective collaborators** (G-1:4)

  In a variety of settings with diverse groups of people

- **Experiential learners**

  Able to integrate curriculum with real-life situations (G-1: 5, G-2: 3,7)

- **Responsible global citizens**

  Taking responsible roles in a global society promoting the quality of life (G-1: 6)

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

Of considerable interest was 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.

The impact of these conversations on the thinking of the participants was evident in their *Daily Reflections*. For example, from the initial training session came the following quotations:

> The training session in August was an interesting experience for me. It presented some very exciting prospects, and at the same time left me with quite a few questions. It almost overwhelmed me. So much was presented, and I envision so much to be done that I have asked myself whether I can do justice to the program in my role.
>
> The four of us have come away from the three day meeting with loads of material to read and review. We are working with portfolios to a small extent in our SIA programs at the K-3 levels. The idea of a portfolio assessment in the area of science and math will be a complete change from our present assessment procedures. We hope that this process will be a step-by-step process with many opportunities for question and answer sessions...We all feel that this change from our present types of student assessment to the portfolio assessment process will be a long term project...This portfolio assessment will need to break through many paradigms before being completely accepted as a replacement assessment instrument.
>
> I was a little apprehensive about the task at hand...I did feel that there may have been a little too much data thrown at us in a short period of time, but I realize that it was the time factor itself that caused this problem. Then again, this may have been just because I was not as prepared and knowledgeable as I would have liked to have been. I was not very comfortable developing goals for the project.

Note the enthusiasm and the caution of these participants. Later on, the comments recorded on the *Daily Reflections* became more enthusiastic and more critical. For example, consider the following:

> This was a very useful meeting! I've been hungry to know where other teams are "coming from" and what they're up to. It was very informative to share (useful at many levels!). It's helped us to clarify our own thinking and we feel pretty good about where we are right now...I'm still personally needing a science/math specialist for a reality check. I'm worrying that math modes of thinking/ communicating, etc. are possibly being biased by the heavy language arts orientation.
>
> Could it ever be more true than the statement, You teach what you assess and you assess what you teach...I think that my next big breakthrough will be when I understand and can use "standards" and a scoring system. And I still have BIG reservations about the time restraints/feasibility of all of this in classrooms/schools as they are now structured.

These comments are representative of the group of twenty-four teachers. Comments of the project staff became equally more focused as the project evolved. The sense of what a wonderful opportunity this was to be able to conduct joint research in assessment with classroom teachers from six different school systems—with time to think supported by the National Science

Foundation—was treasured. In addition, the project staff was convinced early in the project that important assessment models would be produced.

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," and so on. 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 3.4.

With these "evidentiary behaviors" as focal points, the school teams were challenged to develop documentation strategies[2] 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.

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

- What were they trying to describe and how?
- What were they trying to document and how?
- What were they trying to model and how?
- 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.[3]

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 3.5, these problems have been listed along with "solutions" tried during the course of the project.

[2] 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."

[3] Cronbach, L. J. (1970). *Essentials of psychological testing.* New York: Harper & Row.

Table 3.4 Evidencs of Effective Collaborators

To develop students who are *Reflective Thinkers* and *Self-evaluators*: • knows his/her learning style, strengths, and weaknesses • knows how to use the identified strengths/weaknesses 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 *Creative* and *Strategic Thinkers*: • 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

Table 3.4 continued on next page

Table 3.4 (*continued*)

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 *Effective Collaborators*: • 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 *Responsible Global Citizens*: • 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 • participates in service activities • participates in the democratic process • identifies values; demonstrates a responsible course of action

Table 3.5 Problems and Solutions

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 revealed, it was essential to clarify, provide specific examples, and to directly model the desired behavior. To support a variety of knowledge during the presentations, we 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 encouraged discussion, reviewed the *Daily Reflections* for the purpose of raising discussion points, and have encouraged informal contact over the telephone or through letters, faxes, and so on.

Relative to "Group Dynamics," a major obstacle was removed when the liaison responsibility for one team was switched from an administrator to team teachers. Interestingly enough, the teachers had not experienced any negative consequences and continued to have rhetorical support and no real interference. However, it was 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 demonstrated extraordinary commitment.

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

As the project staff reflected on the conversations occurring during the large-group meetings, the following shifts were documented early in the project (see Table 3.6). These shifts in paradigms continued to be evident through the end of the project.

Table 3.6 Paradigm Shifts Among Participants

TIME LINE	
From August, 1992 ——————→	**To January, 1993**
Less reflective	More reflective
Narrow perspective	Broader perspective
Simplistic understanding	Complex understanding
Had not been influenced	Had been influenced
Simplistic definition of innovative assessment	"Rich" definition of innovative assessment

Finally, relative to "Frustration," this project confirmed in the minds of the project staff that defining, describing, and implementing portfolio assessment (or perhaps any type of innovative assessment system) may 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.

Like the comments recorded in the *Daily Reflections*, the comments in Table 3.7 indicate both frustrations and the resolution of these frustrations.

Table 3.7 Comments from *Daily Reflections*

"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 removed or lessened the problems remains 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 ended. Some lies in the

personal shifts made by the project partners. And, there was evidence[4] 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.

[4]Extracts from *Daily Reflections*

4

What Worked, What Worked Well, and What Didn't Work at all!

As with traditional assessment work, evidence of utility and meaningfulness is in the examination of student responses. Even the best assessments from a theoretical perspective fall far short if the evidence obtained from students is not appropriate, accurate, timely, informative, or useful. Throughout the development phase of this project, evidence was gathered from a variety of sources that guided the revision process in important ways. Considering the evidence and its use in refining the assessment model and portfolio entries clarified for us the critical role of using theory, student responses, and critical expert judgment in iterative ways to enhance the quality and utility of the assessments. In this study, early interaction between task and student lead to a variety of actions, from task elimination to task revision or task editing. In all cases, however, the process supported the continued clarification for the teachers of what was important to teach and what was important for students to know, understand, and be able to do in science or mathematics.

Each school team was challenged to author an assessment entry for the science and mathematics portfolio, including the ancillary documents (scoring guides, teacher directions, etc.). The deadline for first draft materials was driven by the opportunity to field-test the assessments during spring of 1994. Each team was further challenged to craft an assessment tool that would address two or more of the goals in either science or mathematics, or in some interdisciplinary context. The particular content to be tested was not specified; instead, each school team wrestled with whether or not particular and specific content was part of the assessment definition or whether assessments could literally be "content free." The distinction made by the project staff was that the study of some goals in some content may indeed lead to an assessment structure and format that would be unique to that content. Therefore the assessment would likely only be appropriate for specific content. On the other hand, for some goals and some content, there might be many units of content that would result in equivalent learning. Thus, the school teams had considerable flexibility in setting their own course of development.

For some teams, the assessment ideas were derived directly from favorite instructional practices. For others, the assessment ideas were derived from new literature, and still for others, the ideas derived from almost incidental thoughts or experiences. Watching the assessments unfold over several months, it became clear that there is no one correct and exclusive path to assessment idea formation. If anything, the work of crafting meaningful portfolio entries is much more difficult and problematic than the work of generating traditional short-answer or multiple-choice questions.

It also became clear that the distinction between learning activity and assessment activity was foggy. We often had to revisit the goals to remind ourselves that indeed there was no specific content referenced in any of the goal statements—because in elementary and middle grades science, this group of teachers was not prepared to say "all students must study and learn _____." Instead, the influence of the *NCTM Curriculum and Evaluation Standards for School Mathematics* (1989), *Science for All Americans* (1989, 1990), and *Benchmarks for Science Literacy* (1993) on these participants was to empower teachers to select the appropriate instructional stimulus given the needs and interests of specific groups of students rather than to teach a particular content area because that was what was on the formal school agenda. In making these choices, of course, the parallel of empowerment is responsibility, and the teachers had to face the reality that as they made choices of what content to teach, how to teach the content, and when to teach specific content, they were also going to examine whether or not their choices had been appropriate through the assessment evidence.

Assessment Development

The documentation strategies presented are similar; the majority of teachers wanted to conduct extensive interviews of all students. Teachers were convinced that only by individual one-on-one questioning could they really find out what their students thought, knew, and could do. Other suggested documentation strategies included the use of logs and laboratory reports. There were no innovative strategies proposed initially.

Prompted to continue to think about documentation strategies in new and different ways, the project staff pushed the school teams to go back to the drawing boards and to think about the evidence they wanted to elicit from students about the goals and about specific content standards from the reference materials.

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

alternative 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 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 that 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. The emphasis on variation, then, 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.

In addition to preparing the final versions of the documentation strategies for field testing, the research partners were also challenged to develop "first tries" at a scoring rubric to be used in informing the students and parents of the valued evidence. Research partners were also 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 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 4.1).

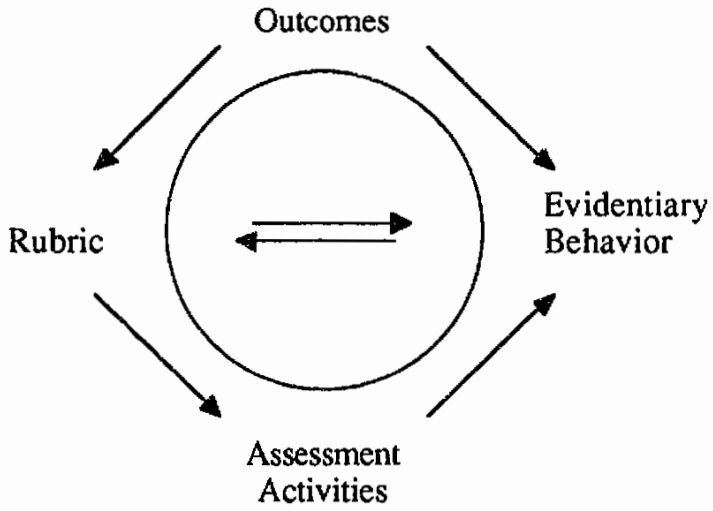


Figure 4.1 Assessment Development Cycle

The teacher-as-stakeholder dimension has been explored through the development of nomothetic standards and judgments. For some research partners, work began on moving into other dimentsions, with consideration of parents, students, and evaluators as stakeholders. However, in general, it is accurate to report that the work progressed slowly into the other dimensions. It is also accurate to report that consensus within these other dimensions was less of

an issue than was the design of relevant and relatively context-free assessments. In fact, the teachers as advocates of this portfolio paradigm had relatively little resistance if their presentations were couched in general terms of having the assessment format appropriate for the evidence of student learning that needed to be documented.

Consistent with the perspective of a vocal minority within 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 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 enable teachers to facilitate the involvement of students in the selection of the stimulus itself. However, it is 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 was 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 the burden to administer for both student and interviewer and the burden for documenting the interviews, including the consequent burden of summarizing or scoring the documentation.

The words from Einstein and Snow were used to remind the research partners of the realities of shifting to new paradigms—that even the new paradigm may not do all that is desired and that even the optimal paradigm will negatively impact someone or something:

> "Not everything that counts can be counted,
> and not everything that can be counted counts," and
> "No matter how you try to make instruction
> better for someone, you will make it worse for someone else."[1]

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, tried 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, a picture that is dependent on information that 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 that will eventually reflect certain limitations.

A second issue of concern was the absence of evidentiary behaviors for any goal that articulates student learning in terms of the knowledge and processes of science or mathematics. Although the goals embraced by this project were derived from the philosophy underlying the *NCTM Standards*

[1]Richard Snow (1989). *Abilities, motivation, and methodology*, The Minnesota Symposium on Learning and Individual Differences.

(1989) and *Science for All Americans* (1989), the direct and explicit linkages were missing. Immediate work began on expanding the evidentiary behaviors to articulate the explicit linkages.

Because the focus of this project was to expand the traditional paradigm for portfolio assessment to one that supported aggregation over time and student groups, the priority for assessment task development was to create evidence-eliciting tasks that addressed the project goals, capture evidence of important learning in the disciplines as defined by the standards documents, and assemble a collection of assessment tasks that in their totality represented the four documentation dimensions, the seven goals, and relevant content. The assessment tasks presented in the following chapters approach this ideal.

5

Letter Writing

This chapter includes the *Letter Writing* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form, and readers are invited to reprint and use them.

This chapter begins the story of assessment task design, implementation, and scoring with the most familiar looking assessment strategy called *Letter Writing*. It is familiar because it looks like an instructional activity used in language arts in the elementary and middle grades. It resembles approaches used in the direct assessment of writing in large-scale assessment programs. In addition to the four dimensions noted in Chapter 4—time, content complexity, stimulus complexity, and response complexity—the school teams were testing their creative talents as they expanded upon instructional activities (learning activities) and turned them into assessment activities by enhancing their complexity, their richness, and the amount of information yielded.

Relative to the dimensions of *time, content complexity, stimulus complexity, and response complexity*, it scores as follows:

Attribute	Score
Instructional Time	3 to 6 weeks, typical unit
Content Complexity	Variable
Stimulus Complexity	Simple
Response Complexity	Complex

Development History

Letter Writing was developed by the Richmond County school team. This strategy of collecting evidence from students was one that they typically used in language arts. The strategy addressed the primary project goal of "effective communicator" and it enabled each teacher to use this assessment for an instructional unit that they thought was most appropriate given their students. As such, *Letter Writing* is a content free assessment strategy. It is not, however, void of content. This is an important distinction; we believe that it is possible to construct assessment formats that are effective in eliciting important content without specifying it. For example, *Letter Writing* as an assessment strategy may be equally as useful in capturing evidence about what students think, know, and can do with fractions and ratios as with patterns or geometry. Likewise, in science, it may be as useful in eliciting evidence of learning following a unit on buoyancy as one on weather or molecules. The critical feature in these types of assessments is whether or not the scoring guides enable one to report information on specific content even if content is not specified by the assessment activity. We think that *Letter Writing* makes an important first step in this direction.

As you review the Teacher Instructions and the Student Worksheet, remember that this assessment activity was designed to be one of eight entries in a student's science and mathematics portfolio. Around these eight tasks that define the structured core for the assessment portfolio would be additional unique and idiosyncratic pieces of evidence selected by the student and/or the teacher for various reasons. In toto, the assessment portfolio would support aggregation based on the structured core entries and would also support individual pictures of student growth and development based on the structured core entries plus the individualistic entries of choice.

Teacher Instructions

LETTER WRITING ASSESSMENT

Teacher Instructions: Letter Writing

Overview

Students explain, in the form of a letter to a specific audience, what they have learned from a particular unit of study. Additionally, students are to include requirements assigned by the teacher. These idiosyncratic requirements are: (1) appropriate for the unit of study, and (2) not already evaluated in the rubric.

Purpose

The task was designed to evaluate students'

1) Understanding of the topic.
2) Use of appropriate vocabulary.
3) Ability to incorporate knowledge with other learning and/or use the knowledge in their everyday lives.
4) Ability to demonstrate reflective or strategic thinking.
5) Ability to communicate effectively in written form.

Planning

√ **As you plan the letter writing assessment activity, the following should be considered:**

1) This letter writing assessment activity is designed to challenge students to write about a two- or three-day unit of instruction which they have just completed. This unit may be in mathematics or in science. It may also focus on interdisciplinary mathematics and science or either mathematics or science in conjunction with some other content area.

√ **Read the rubric before selecting the experiments; the knowledge will help in guiding your selection.**

2) Language mechanics, such as spelling, punctuation, and grammar, should be handled in such a way that the students are not deterred from taking risks in the writing task.

3) Assigned requirements (those determined by the teacher) should be discussed with the students prior to beginning the assessment activity. These writing assignments must be able to be evaluated as either present or absent since the rubric does not accommodate partial compliance.[1]

[1] The assigned requirements will differ with the specific activity and must be stipulated by the teacher in advance of the assessment. There should be at least three but not more than five assigned requirements. Assigned requirements can include such items as citing examples, expressing opinions, supporting arguments, suggesting alternative procedures or solutions, making writing consistent with the relationship between the writer and the addressee, and using standard English throughout the letter

 Project funded by The National Science Foundation.

4) Information about the rubric must be shared with the students. The students should know that their work will be evaluated in terms of how each demonstrates that he or she:

- Understands and can explain the activity.
- Uses appropriate vocabulary.
- Incorporates assigned requirements.
- Incorporates knowledge with other learning and/or uses the knowledge in everyday life.
- Thinks reflectively about personal performance and/or the information presented and/or the value of the information.

√ **The attached information sheet must be completed for scoring purposes. It is also useful as a "check-list" to evaluate the appropriateness of the selected topic and experiments for this task.**

Teacher _____ **School** _____

Information for Scoring
Letter Writing

This information is necessary for scoring student responses and should be completed before the assessment is administered. Answer each question as completely as possible.

☞ *Describe the activity or unit of study (be specific):*

☞ *What vocabulary was presented in the lesson that students could be expected to use?*

☞ *What concepts or essential elements should be included in an explanation of the activity or in response to what was learned? It is imperative that the concepts/elements listed were clearly presented in, or a focus of the instruction.*

☞ *What were the additional requirements that you assigned for inclusion in the letter? Make sure you are specific about what is expected.*

Letter Writing: *Information for Scoring*
 Project funded by The National Science Foundation.

Student Worksheet

LETTER WRITING ASSESSMENT

Letter Writing
Student Response Worksheet

Student Name: _____

School Name: _____

System Name: _____

Teacher Name: _____

Grade: _____ Date: _____

Directions:

Your task is to write a letter to someone who is important to you about the activity that you just completed. The person to whom you write can be your parent, your teacher, your best friend, the principal, or anyone else you choose. In the letter:

- Explain the activity.
- Use the vocabulary that was presented.
- Make sure your explanation is organized and clear.
- Explain what you have learned.
- Explain how you could use the information in other situations.
- Include any questions that you have that were not answered during the activity or thoughts you have about your understanding of this topic.

Include EVERYTHING your teacher has asked for in this letter - list them below:

Letter Writing: *Student Worksheet*

Student Response Worksheet: Letter Writing

Second Page

Letter Writing: *Student Worksheet*

Following is the scoring guide or rubric for this assessment activity. The scoring rubric for Letter Writing is holistic rather than analytic. As such, the student products -- letters -- are to be judged with a single, overall statement of the quality of the response. The decision to use a holistic scoring guide rather than an analytic one was made by the development team and was a result of their understanding of the information needs of the stakeholders as defined in the portfolio assessment model under study. Other assessment strategies described later in this chapter use analytic scoring guides. Analytic scoring guides yield separate scores on specific qualities or elements of the product being judged.

The number of score categories for Letter Writing (0 through 5) was determined based on the review of the advice of experts in scoring student work merged with the information needs of the model stakeholders as well as try-outs of the task to determine the extent to which different levels of performance were demonstrated. Scoring guides in the direct assessment of writing -- our most well-researched performance assessment field -- typically include scoring guides with four to six score points. Some may include an additional category of "no response" or "off task" to indicate that a particular student essentially made no effort to respond to the provided stimulus material. The range of performance levels needs to be meaningful and it needs to virtually reflect student work. Thus, the development teams were challenged to define performance levels that would literally subdivide the sample of student work so that there was minimal disagreement among trained professional educators about the score assigned. The labels for the performance levels were selected by the development teams as meaningful descriptors.

Note that the interpretation of each scoring guide relies heavily on the exemplars provided during training for the scoring.

Scoring Guide Letter Writing		
5	Exceptional	Explanation is well-developed, clear, and elaborated, without content errors. There is extensive use of presented terms with demonstrations of precision and command of terms. All assigned requirements are present. Explanation extends beyond requirements of the task in the following ways: • Information given includes reference to other circumstances/content/topic, prior knowledge/interests, connections/application to everyday life • Explanation includes evidence that student reflects on learning and the value of learning and monitoring of thinking about the topic/activity/explanation.
4	Very Good	Explanation is clear and developed. There is evidence that concepts presented are understood. Appropriate use of all presented terms. ALL assigned requirements are present. Explanation extends beyond requirements of the task in the following ways: • Information given includes reference to other circumstances/content/topic, prior knowledge/interests, connections/application to everyday life *OR* • Explanation includes evidence that student reflects on learning and the value of learning and monitoring of thinking about the topic/activity/explanation.
3	Satisfactory	Explanation is clear but may be unevenly organized or loosely developed. There is evidence that the concepts presented are understood, but there may be minor errors. There is appropriate use of most presented terms. Most of the assigned requirements are present.
2	Limited	Explanation is poorly developed. There is some evidence that the concepts presented are understood. Some presented terms are used appropriately, or presented terms are used but with some errors. Some of the assigned requirements are present.
1	Minimal	Explanation of topic/activity is unclear or poorly developed. There is little evidence that the concepts presented are understood. Presented terms are not used, or few presented terms are used correctly. At least one of the assigned requirements is present.
0	Off Task	Student did not attempt task or did not follow directions. Concepts presented were not understood.

Letter Writing: *Scoring Guide*
 Project funded by The National Science Foundation.

Because *Letter Writing* is essentially a content free assessment tool, it is important to scrutinize whether or not the scoring guide or rubric is sufficiently sensitive and detailed to enable one to report out information on content. This becomes increasingly important as the criticism mounts about the failure of new forms of assessment to provide information about whether or not students are learning *important content.* In fact, we may literally be shooting ourselves in the foot if we fail to address this important information request from parents, administrators, and education policy-makers. It may be that this assessment package requires redefinition of the performance levels in order to meet the information needs of those who want specific content scores. However, this can be accommodated by revising the scoring guide. We invite the reader to think through this issue carefully and to modify the scoring guide as needed.

Note also the important pages detailing the specific instructional stimulus to be selected by individual teachers for this assessment (pages 36-37). Without this level of detail to inform scorers about what eligible and appropriate content is, no reliable or valid scoring of student work could proceed. This information is not necessary in assessments that prescribe content. It is, however, critical in those assessments that are content free.

Examples of Scored Student Work

The five exemplars selected for inclusion here were used to train teachers to make reliable and comparable judgments about student work on the *Letter Writing* task were obtained from a small-scale try-out. They are included here as typed italicized text but in the training materials used they were literally copies of the handwritten letters. Each exemplar is notated with a score and a brief rationale. The identification of each response is a coded student number to preserve anonymity. The scoring guide for *Letter Writing* is a five-point[2] holistic scale (Exceptional [5], Very Good [4], Satisfactory [3], Limited [2], Minimal [1], and Off Task [0]).

[2]This is considered a five-point scale because the off task responses are not even considered in terms of designating quality.

Example 1

Student D0073 was instructed to write a three-paragraph letter that would persuade a friend, judge, or teacher why the project had merit. The science project topic was each student's choice. This response was given a score of 5 because it met all requirements for exceptional (well developed, uses terms— demonstrates knowledge of types of eclipses, includes all requirements and explanation goes beyond the requirements of the task).

Dear *Mom and Dad,*

I plan on entering a project on solar eclipses in the (school name) Science Fair. I chose the solar eclipse because I have always thought they were beautiful and unique. What I cant belive is that the moon blocks out the sun from the earth!

I have learned from all of my research that the shadow of the moon moves about 2,000 miles per house acros: the earth. I also learned that there are three types of solar eclipses, which are the total eclipse, the annular eclipse, and the partial eclipse, The total eclipse occurs when the moon blocks out the sun completely, the annular occurs when the moon is at its farthest point from the earth and completely blocks out the sun, and the partial sclipse occurs when the moon only blocks out part of the sun.

I think my science project is very original because no one else in the class has done or even thought about doing a solar eclipse project. I don't know if my project is good enought to get frist, second, or even third prize because I have seen some pretty wonderful and unique projects, but there still that chance. I really don't care about having a fancy ribbon, all I care about is doing my best. These fancy ribbons are nice, but there not everything.

When I gro up I want to be a Marine Biologist, but I will always be interested it all sciences.

Your son,

(name)

(included graphic of full eclipse as well)

Example 2

Student R0104 had just completed a two-day unit on Landforms. The landforms presented were mountain, plain, and plateau. Each landform was defined in the unit. The students practiced drawing each type of landform. During the unit students also worked in groups. Each group was given an atlas in which they were to find examples of each landform throughout the world. For the activity, the students were told to define each landform, give any facts about the landform they had learned during the unit. They were to draw and label an example of each landform. This response was given a score of 4 because it extends but does not reflect on the content. Both of these features are needed for a score of 5.

Dear *Clemetine,*

We have been talking about landforms. I will start with Plateaus! Plateaus are highland that has low relief and it rises high above sea level flot on top. You can find a plateau in Ethiopia. Plains! Plans are land close to sea level called lowlands. Plains are lowlands that are faily flat. One plain area in U.S. lies on eastern Coast. Massachusetts to Gulf of Mexico, Covers much of the Mlddle of North America Called the Great Plains. Mountains: Mountains are the tallest landforms. The highest peak is Mount Everest it is in Asia, U.S. Mount McKinley is in Alaska, always in groups called ranges, sereral ranges = a chain, Rocky Mountains, at leats 600 meters above the Surrounding land.

(included a picture of mountains with a flag)

Example 3

Student R0121 had just completed the two-day unit on landforms detailed in the description of Student D0104. This response was given a score of 3 because it did not contain a drawing or picture and therefore could not earn a score above a 3; all other requirements for a 3 met.

Dear *Mrs. Pearson*

Today, we studied about landforms. A landform is a shape of the land, such as a mountain, plain, or plateau.

Mountins are the highest landford on Earth. Mountains grouped together is called a mountain range. Mountain ranges grouped together is called a chain of mountains. Mt. Everest is the highest peak in Asia and Earth. The Rocky Mountins are located on the west coast.

Plateaus are like flat land. It is highland with low relief. It might have grass on top. They have no peak on top.

Plains are flat land. It might have a few hills. The middle west is called The Great Plains.

In class, we worked in groups and used an Atlas to look at landforms drew pictures.

Sincerely -

Example 4

Student F0064 had just completed a two-week review of multiplication. The original instruction had been during the first quarter of school using the text, the curriculum guide, worksheets, manipulatives, etc. This was simply review, so most of the students were comfortable with it. After these two weeks of review, the letter writing assessment was embedded in the day's lesson. The students were directed to write a letter explaining what multiplication is. They were told that they could write to their parents or any other relative or to a friend. After the letters were written, those who wished to do so were allowed to share theirs with the class. This response was given a score of 2 because it seems to be a basic understanding but explanation is not developed, terms not used, "work" does not enhance explanation requirements for a 3.

Dear *Shanika*

You do not no how to Multiply bu um I am going to show you how. Play like I ask you what's 5 x 8 -. Play like u have five cars and you count them eight times and that the way you multiply.

Love your big bother

Look on back

1 x 2 = 2
1 x 3 = 3
1 x 4 = 4
1 x 5 = 5
1 x 6 = 6
1 x 7 = 7
1 x 8 = 8
1 x 9 = 9
1 x 10 = 10
1 x 11 = 11
1 x 12 = 12

	x 5
1	5
2	10
3	15
4	20
5	25
6	30
7	35

Example 5

Student G0121 had just completed a two-week unit on plants. The concepts and vocabulary included in this unit were that plants made their own food using photosynthesis, a process in which light energy, carbon dioxide, and water change into food and oxygen in a chloroplast cell. The chloroplast cell has chlorophyll, which makes the leaves of the plant green. To use this food, the plant uses respiration, in which food and oxygen are changed into carbon dioxide, energy, and water. The plant receives sunlight through its leaves. Water and minerals are transported from the roots to the leaves through tubes called xylem. Oxygen and carbon dioxide enter the leaves through tiny holes called stomata. Two experiments were conducted. First, the students attempted to find out what would happen if the stomata of a plant were blocked, or if air was limited. This activity was done in groups of 5 or 6. Each group had three bedding plants. One was the control plant. One plant was coated with petroleum jelly. One plant was enclosed in a plastic bag. The plants were observed for one week. In the second experiment, students were to find the xylem and observe them transporting liquid. Each group had celery and a cup of water colored with red food coloring. The students were to cut off a piece of the celery from the bottom of the stalk and insert it into the colored water. After waiting about 15 minutes, they took the celery out of the water and cut the celery to see if they could find the xylem.

Student G0121's response was given a score of 1 because it was minimal. It did not receive a 0 because there was some understanding of the experiment.

Dear *Mom and Dad*
I don't know nothing about plants. But if you breack a pice of celery in half you can see the veins.

Each of the teacher-selected units seems appropriate for the assessment task. As evident, *Letter Writing* has the capability of eliciting a wide range of evidence from students.

6

Science Observation

This chapter includes the *Science Observation* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form and readers are encouraged to reprint and use the materials.

Chapter 6 continues the presentation of tasks with *Science Observation*. Just as *Letter Writing* this seems like an instructional activity that could be used in elementary science classes. It is, however, somewhat unusual for an observation activity to become a systematic assessment. This feature makes *Science Observation* unique. In addition, *Science Observation* provides the development platform for *Toys in Space* presented in detail in chapter 13.

The second assessment entry for the assessment portfolio developed was *Science Observation*. This task derived directly from the work of Ellen Doris. In fact, both the notion of communicating about science through drawing and the value added to that by presenting the drawing to peers for review comes from Doris' book, *Doing What Scientists Do*[1]. This book was brought to one of the early development sessions by one of the participating teachers.

Doris described science as a process of inquiry and investigation—a way of thinking and acting— and through her book she emphasizes that science is not only a body of knowledge but a habit of mind. Thus, the connection between her work and that of *Science for All Americans* (1989, 1990) and some of the strands in the *NCTM Standards* was clear. Like *Letter Writing*, *Science Observation* is an assessment strategy that looks like a common, everyday activity in many content areas. However, the essence of the work is that students must systematically examine what they see in their world and that they must be able to communicate their observations accurately and clearly to others.

Relative to the dimensions of time, content complexity, stimulus complexity, and response complexity, the task is scored as follows:

Attribute	**Score**
Instructional Time	Variable
Content Complexity	Variable
Stimulus Complexity	Variable
Response Complexity	Variable

Compared with the dimensions of time, content complexity, stimulus complexity, and response complexity present in *Letter Writing*, *Science Observation* is less constrained. Therefore, the character of the evidence of student learning elicited by this task is very much at the control of the teacher. And, what we quickly learned through field trials was that some content simply did not present an appropriate stimulus for this assessment. Quite simply, in order for there to be meaningful evidence documented by the *Science Observation* task, the science content had to be appropriately "observed and described" through drawing. We found that motion in particular was not an appropriate content stimulus for this assessment; it was simply too difficult for students in elementary and middle grades to capture motion effectively through line drawings on paper. On the other hand, science content that could be frozen in time as with a figurative "snapshot" was appropriately documented through this process. Clearly, the content eligible for this assessment task greatly exceeds the pool of inappropriate content. The case simply needs to be emphasized that as teachers make decisions about which assessment strategies to use, they must consider which will most effectively and efficiently capture evidence about what students think, know, and can do in light of content.

[1]Doris, E. (1991). *Doing what scientists do—children learn to investigate their world.* Portsmouth, New Hampshire: Heinemann.

Science Observation was developed by the Clarke County school team. The strategy addressed the primary project goals of "effective communicators," "effective collaborators", and "reflective thinkers and self-evaluators." The structure of the task is essentially that a student draw something observed. This drawing is shared with peers to determine whether it communicates effectively what is intended from the perspective of the peers. This feedback is then used by the student to refine and redraw the object/phenomenon of interest.

Like *Letter Writing*, *Science Observation* is a content-free assessment strategy. It is not, however, void of content. And, again a critical feature in these types of assessments is whether or not the scoring guides enable one to report information on specific content even if content is not specified by the assessment activity.

Teacher Instructions

SCIENCE OBSERVATION ASSESSMENT

Teacher Instructions: Science Observation

Overview

Students will be asked to demonstrate their ability to make detailed, accurate scientific observations through drawing and writing. Students will view an object and produce a scientific drawing of their observation. Additionally, students will be asked to describe the object in writing.

Purpose

The task was designed to evaluate students'
1) Observation skills as evidenced in their scientific drawings
2) Ability to effectively describe the objects observed

Planning

√ **In the plan for the observation task, the following should be considered:**

1) The object(s) to be observed must be complex enough to note multiple relevant features but not so complex that the student does not know what to focus on. For example, the school playground is too expansive an area to observe for this task; a simple potted plant or a microscopic view of a drop of pond water is more manageable.

√ **Read the rubric before selecting the object; the rubric will help in guiding your selection.**

2) Students should have received instruction on scientific observations and ample practice in drawing and describing other objects prior to the assessment.

3) Share the rubric with the students before they begin the assessment. The rubric can serve as a self-evaluation tool during instruction so that students know what is valued in this activity.

4) The teacher should provide appropriate materials for drawing, considering the students' age and the object(s) to be drawn.

√ **The attached information sheet *must* be completed for scoring purposes. It is also useful as a "check" to evaluate the appropriateness of the selected object for this assessment task.**

Teacher: _____ School: _____

Information Sheet for Scoring
Science Observation

1. In the space below, provide a detailed "master" drawing of the object(s) students are to draw. The master drawing serves as a template of observations students have the opportunity to make. In the master drawing, note the relevant features and specific details that students might include.

Science Observation: *Information for Scoring*

2. In the space below, provide a written description of the object(s) and the features detailed in the "master" drawing. Make sure you include objective, precise descriptors. Include observations of touch, smell, and taste, if appropriate.

Science Observation: *Information for Scoring*

Student Worksheet

SCIENCE OBSERVATION ASSESSMENT

Science Observation: *Information for Scoring*
 Project funded by The National Science Foundation.

Science Observation
Student Worksheet

Scientist's Name: _____

School Name: _____

System Name: _____

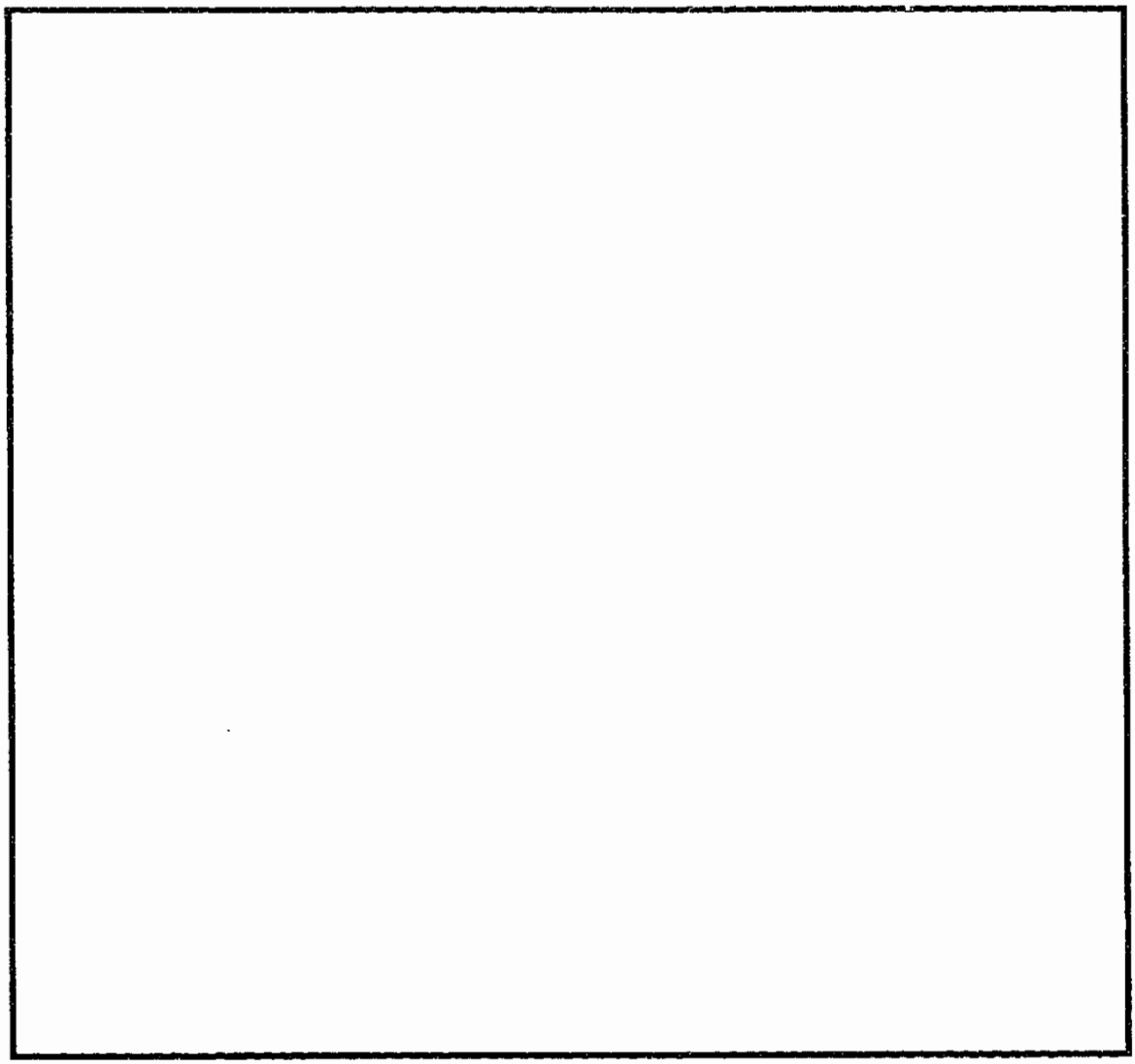
Teacher: _____

Grade: _____ Date: _____

I observed:

FIRST DRAWING

A drawing of what I observed:



A written description of what I observed:

SECOND DRAWING

A second drawing of what I observed:

Write about which drawing is better and why you think so.

As with *Letter Writing*, the development team for *Science Observation* also elected to develop a holistic scoring guide. Likewise, the development team selected a five-point performance scale with an additional "0" score for off task performance.

Scoring Guide Science Observation		
5	**Exceptional**	The drawing includes multiple relevant features with attention to many specific details (shape, size, color, texture, proportions, parts, unique features). Drawing is a clear visual representation; there is accurate use of scale. The written response describes the object(s) observed, noting the multiple relevant features and specific details (exacting, clarifying, unique descriptors). The written description is objective and precise in the use of descriptors and includes those features drawn. Additionally, the description includes those elements observed through senses other than sight
4	**Very Good**	The drawing includes multiple relevant features with attention to some specific details. Drawing is a clear visual representation; there is accurate use of scale. The written response describes the object(s) observed, noting the multiple relevant features and some specific details. Overall, the written description is objective and precise in the use of descriptors and includes those features drawn. Writing is more than a listing.
3	**Satisfactory**	The drawing includes multiple relevant features with some attention to detail; irrelevant features may be included. Drawing is a clear visual representation; there is evidence of an understanding of scale although scale may be imprecise. *and/or* The written response describes the object(s) observed, noting multiple relevant features and some specific details; irrelevant features may be included. Overall, the written description is objective and precise in the use of descriptors and includes those features drawn.
2	**Limited**	The drawing includes few relevant features with little attention to detail. Drawing is identifiable but may be overly simplistic. There is a distortion of scale that demonstrates confusion about proportional relationships. *and/or* There is an attempt to describe the object(s) observed, but with few relevant features or details. Overall, the written description is vague and may include inconsistencies when compared to the drawing.
1	**Minimal**	The drawing includes few relevant features with little attention to detail. Object is difficult to identify. Scale is distorted. *and/or* There is an attempt to describe the object(s) observed, but with few relevant features or details. The written description is vague and may include inconsistencies when compared to the drawing.
0	**Off Task**	No attempt was made or student did not follow directions for task; there is not enough information to score.

Science Observation: *Scoring Guide*
 Project funded by The National Science Foundation.

Examples of Scored Student Work

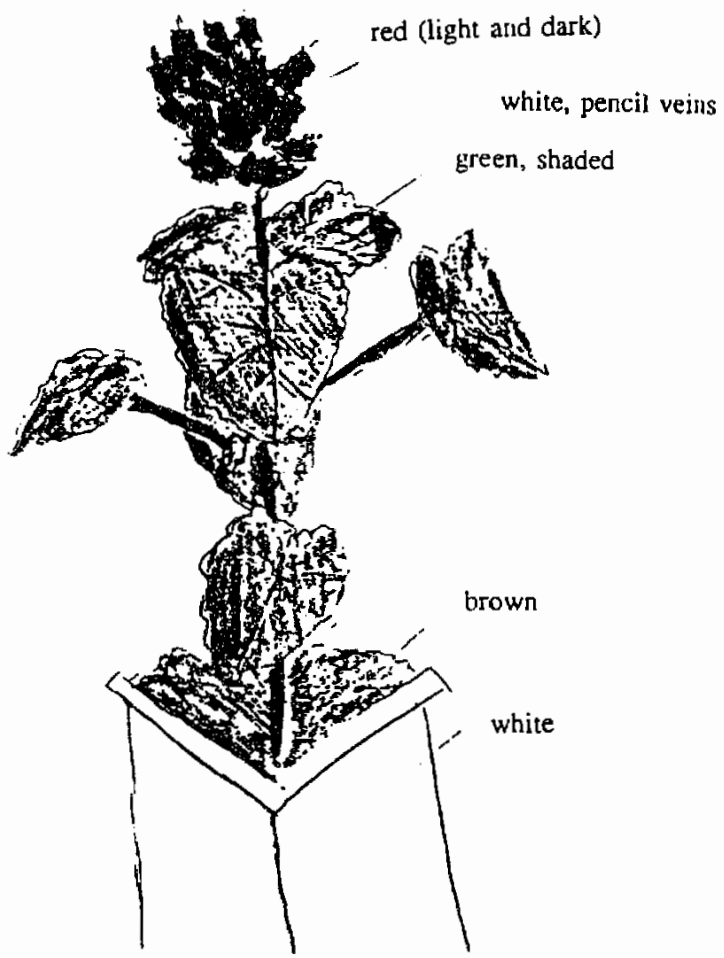
The five exemplars selected for use inclusion here were used to train teachers to make reliable and comparable judgments about student work on the *Science Observation* task were also obtained from a small-scale try-out. They are included here as typed italicized text. The drawings are black and white copies with the colors written in. The identification of each response is a coded student number to preserve anonymity. The scoring guide for *Science Observation*, like that for *Letter Writing*, is a five-point holistic scale (Exceptional [5], Very Good [4], Satisfactory [3], Limited [2], Minimal [1], and Off Task [0]).

Example 1

Student G0113 had studied plant structure. This response earned a score of 5 because the drawing included multiple relevant features (flower, stem, leaves, soil), with attention to detail (two-toned leaves, veins outlines in leaves, flower petals separated, varying leaf sizes). It provided a clear visual representation, including accurate use of scale and perspective. With respect to the writing dimension, this response describes the object, noting multiple relevant features (plant, leaves, stem, flower, contained), with attention to detail (leaves are dark green on one side and light green on the other ...on the green stem are small leaves...up at the top is a flower...flower is red and looks like a bunch of small flowers making a big flower...). In addition, the writing is objective with precise use of descriptors, includes features drawn; evidence that other senses were used (fuzzy leaves...plant smells like a type of herb).

FIRST DRAWING Student G0113

A drawing of what I observed:

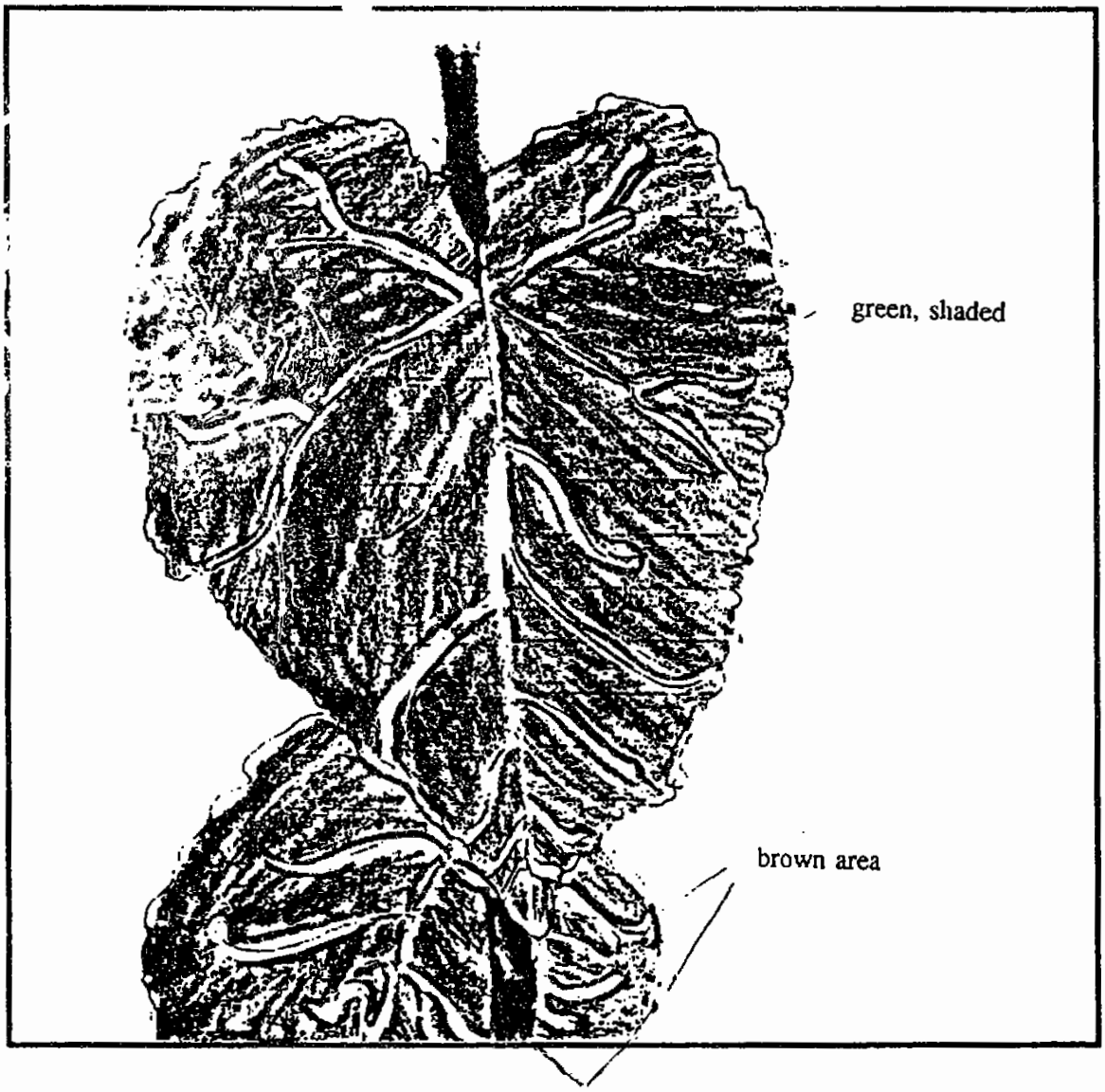


Some things I noticed:

The plant has fuzzy Leaves. The leaves are dark green on onse side and light green on the other. On the green stem there are small leaves. Uptop is a flower. The flower is red. It looks like a bunch of small flowers maing a big flower. The plan smells like a type of herb. The plant is in a white, square, plastic contained filled with dark brown soil. The leaves have little cracks in them. The white contained is in a clear, round, contained. On the clear container is a piece of tapoe that says TEAM 22-20 BEAN SEEDS

SECOND DRAWING Student G0113

A second drawing of what I observed:



Write about which picture is better and why you think so.

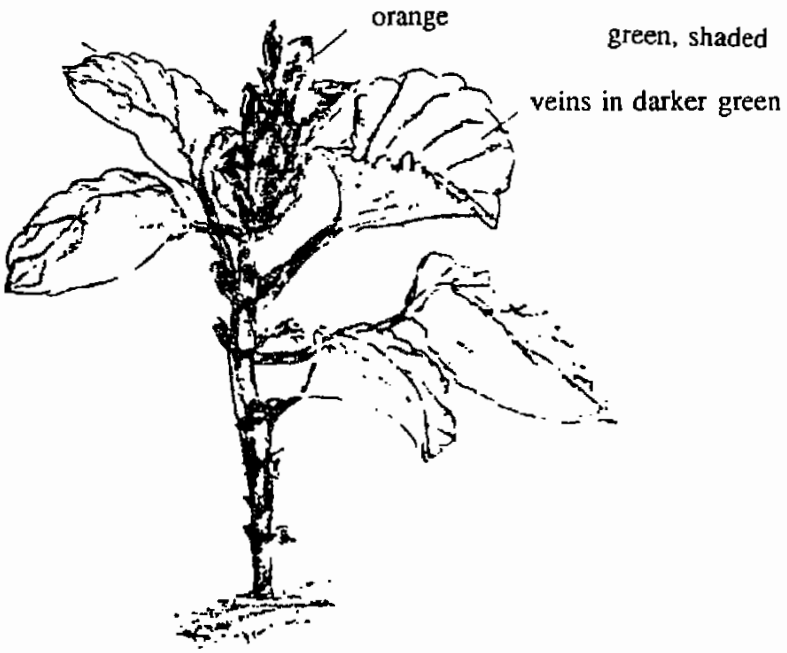
I think my second one was better becaus it had more detail. You call tell more distinktively that it is a leaf on a plant.

Example 2

Student G0103's response earned a score of 4. The drawing evidence included multiple relevant features (flower, stem, leaves, dirt) with attention to detail (veins in leaves, segmented parts of stem, separate petals of flower, coloring is shaded and shows details of features). The writing dimension describes the object notes multiple relevant features (leaves, veins, stem, flower) with some specific details (little purple knot...heart shaped leave...little leaves...big leaves...stem goes big to small); it is objective and precise in the use of descriptors; does not use other senses.

FIRST DRAWING Student G0103

A drawing of what I observed:



Some things I noticed:

I saw a little purple knot. I saw a heart leave shape. I saw some little leaves by the big leaves. I saw lots of veins on the leaves and the stem some part it's big and some part it's small. I saw a little string from the middle of the flower come out.

SECOND DRAWING Student G0103

A second drawing of what I observed:

Write about which picture is better and why you think so.

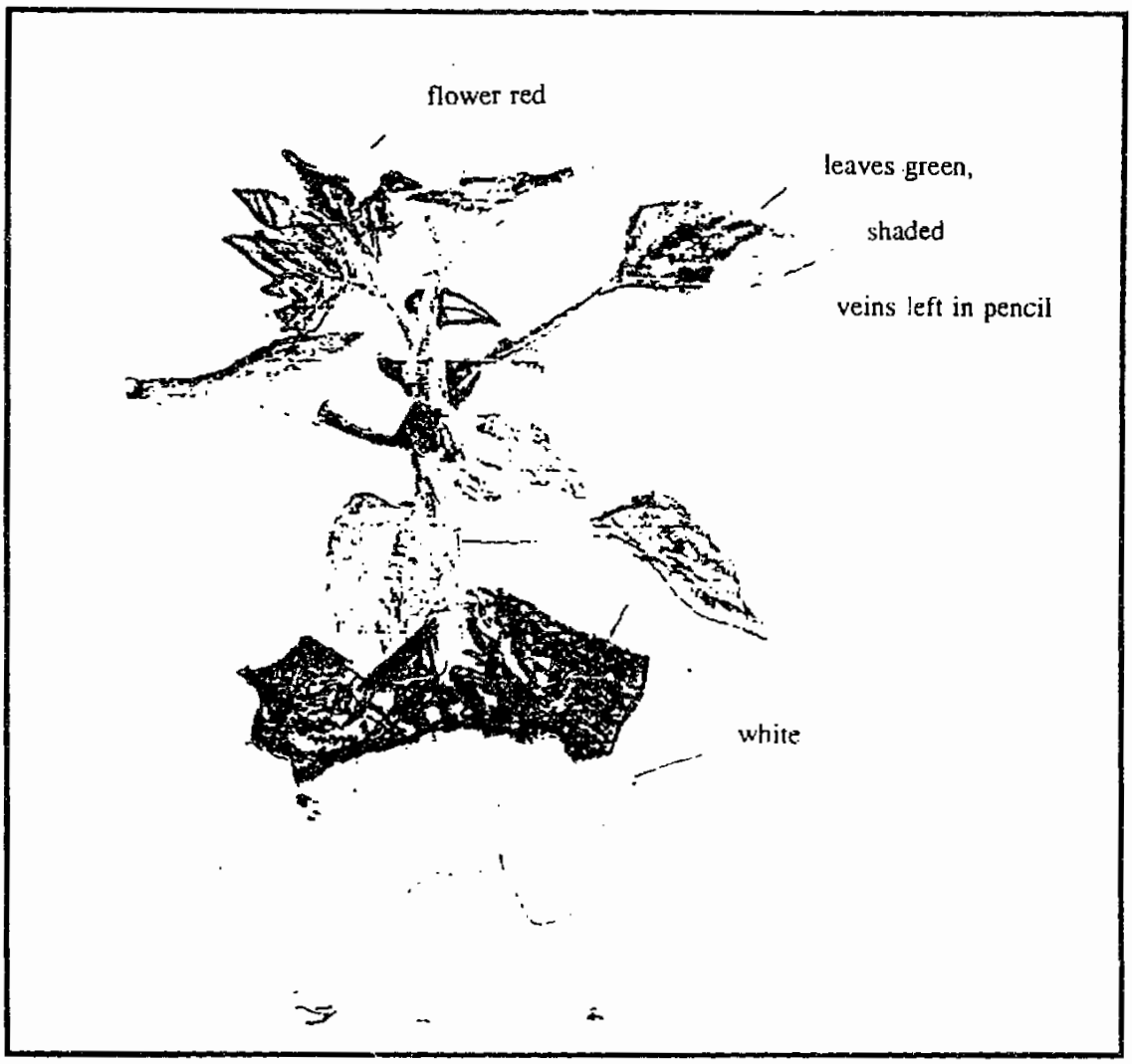
I think I like the first one, because the first one I draw more detail than the second one, also the second was not finish. The first one I draw kind like the real plant and exact size. The second one I have to draw bigger at is it.

Example 3

Student G0105's work earned a score of 3. The drawing evidence includes multiple relevant features with some attention to detail; it is a clear visual representation; evidence of an understanding of scale. With respect to writing, there is minimal description with only a few relevant features of those drawn.

FIRST DRAWING Student G0105

A drawing of what I observed:



Some things I noticed:

I noticed small leaves and little white dirt in the soil. I also noticed some leaves darker than others.

SECOND DRAWING Student G0105

A second drawing of what I observed:

Write about which picture is better and why you think so.

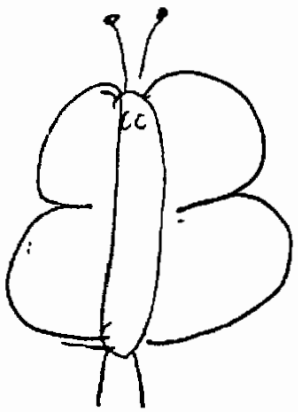
I think the first plant is better because theres more plant to draw.

Example 4

Student R0108's assessment was completed after a unit on butterflies. Thus, the content here differs from that of Student G0113, G0103, and G0105. The response earned a score of 2. With respect to drawing, the first drawing includes few relevant features with little attention to detail. The second drawing is difficult to identify. In the writing sample, there are attempts to describe the object but there are few relevant features mentioned and irrelevant features are included. The description is vague and subjective (the butterflies are...beautiful).

FIRST DRAWING Student R0108

A drawing of what I observed:

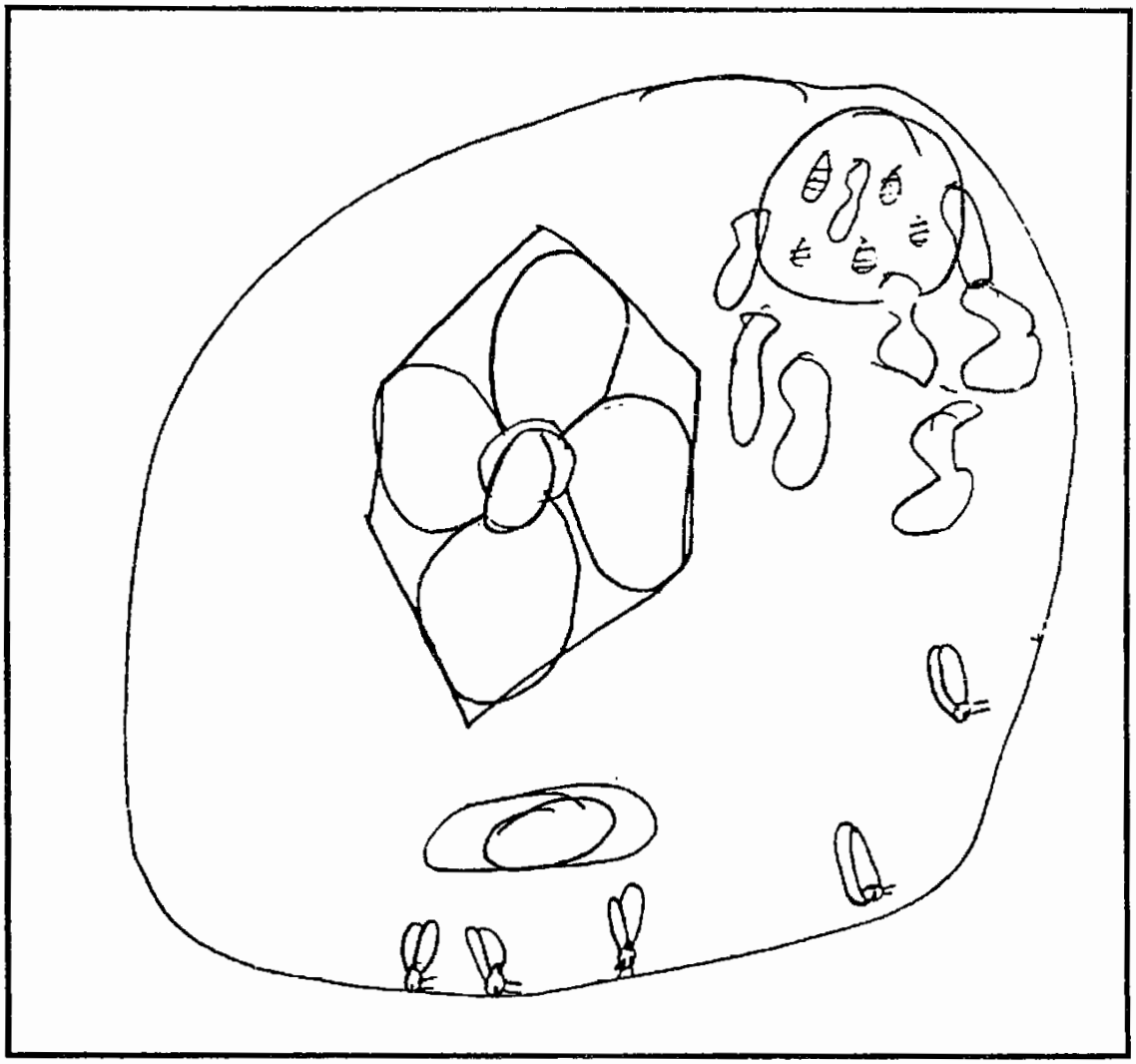


Some things I noticed:

The butterlifes wings are orange, yellow, brown, black. I saw the birds flying. The butterflies are very colorful and beautiful.

SECOND DRAWING Student R0108

A second drawing of what I observed:



Write about which picture is better and why you think so.

I think the second picture looks the best because I had more detail.

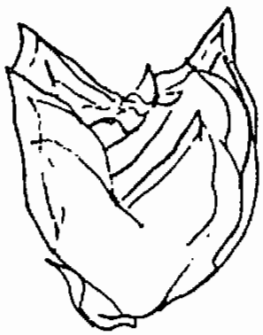
It Just looks BEAUTIFUL!

Example 5

Student G0114 earned a score of 1. The drawing includes few relevant features and the object is difficult to identify. In the writing sample, there is no attempt to describe first drawing. The second description has few relevant features; it is vague and subjective.

FIRST DRAWING Student G0114

A drawing of what I observed:



Some things I noticed:

(no written response)

SECOND DRAWING Student G0114

A second drawing of what I observed:

Write about which picture is better and why you think so.

I like my second one better because it's a close-up veiw of the flower. And it gives you a closer look at the flower. And because I finished it, also because it's colorful.

7

Problem Solving

This chapter includes the *Problem Solving* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form and readers are invited to reprint and use the materials.

Development History

The third assessment entry for the assessment portfolio developed was called the *Problem Solving* assessment. This task derived directly from the teachers' understanding of the scientific method of investigation, and the structure of the *Student Worksheet* looks very much like a lab report form used in many science classrooms. It is, unlike *Letter Writing* or *Science Observation*, particularly limited to the content of science. Mathematics content and process are likely to be evident only as tools for the solution of a problem. It is, however, relatively unspecified in its content requirements with the exception that the content is most appropriately science and that the stimulus situation is a problem to be solved.

Relative to the dimensions of time, content complexity, stimulus complexity, and response complexity, it scores as follows:

Attribute	Score
Instructional Time	Unspecified since the content covered may span an entire year of science study
Content Complexity	Variable
Stimulus Complexity	Variable
Response Complexity	Relatively limited by the implicit paradigm suggested by the Student Worksheet and its resemblance to a standard lab report form

Compared with the dimensions of time, content complexity, stimulus complexity, and response complexity present in either *Letter Writing* or *Science Observation*, *Problem Solving* is more constrained. If the instructional program modeled lab work for the students, this assessment would seem to be much more directly and obviously linked to learning activities and to teaching activities. It is, perhaps, so obviously an appropriate example of documenting student learning that the reader might say "why is this even included as representing new forms of assessment? We have been doing this for years!" And, quite so from that perspective. However, what hasn't been happening for years in most classrooms across the country is that accompanying such systematic ways of collecting and documenting what students think, know, and can do is a scoring guide that specifically and explicitly defines levels of quality work.

The most important lesson to be gained from *Problem Solving* as a model for assessment is that there are indeed many reasonable and appropriate tools for capturing evidence about student learning that can readily be transformed into assessments that are systematic, fair, and credible. But, part of the transformation is the important feature of developing a scoring guide that can be used to systematically examine student work for the purpose of judging levels of quality.

Problem Solving was developed by the Fulton County school team. This assessment strategy addressed all of the project goals of "effective communicators," "effective collaborators", "reflective thinkers and self-evaluators," "creative and strategic thinkers," "experiential learners," "responsible global citizens," and "reflective thinkers and self-evaluators." The mode of communication is writing. However, drawing or physical models were not prohibited.

Like *Letter Writing* and *Science Observation*, *Problem Solving* is also content free assessment strategy. It is not, however, void of content. And, again a critical feature in these types of assessments is whether or not the scoring guides enable one to report information on specific content even if content is not specified by the assessment activity.

Teacher Instructions

PROBLEM SOLVING

ASSESSMENT

Teacher Instructions: Problem Solving

Overview

Students are presented with an open-ended experiment. They are to 1) develop an appropriate hypothesis statement or research question, 2) report the steps that will lead to a solution/answer to the original question, and 3) analyze and report the results/conclusions explaining the scientific principles or concepts learned.

Purpose

The task was designed to evaluate students' performance in problem solving/investigation in science. Students demonstrate, in writing, their

1) Ability to explain the problem/task or state a hypothesis/research question

2) Ability to explain or develop an appropriate plan to address the initial problem/question

3) Ability to reach a logical result/conclusion which relates to the original problem/hypothesis.

Planning

√ **To plan the science context (experiments) for this evaluation, the following must be considered:**

1) The experiment must be open-ended and demonstrate scientific concept(s) appropriate for the science curriculum. Students must be able to conduct the experiments by themselves or in groups. It is important to consider what the experiment demonstrates and give careful thought to the appropriateness of the content for the grade level.

√ **Read the rubric before selecting the experiment; the knowledge will help in guiding your selection.**

2) The task and rubric for Problem Solving allow for variability in the presentation of the experiments. For younger students, you may want to introduce the experiment and ask them to explain or restate the problem or task and to report the steps that were taken to determine the outcome. Older, more experienced students may be able to predict the outcome, stating their own hypothesis, and design the steps/procedures to find the answer. The rubric provides a strategy for evaluating both formats. It is critical, however, that the information supplied by the teacher is reported on the attached information sheet.

3) The scientific concepts demonstrated in the experiments must permit accurate conclusions. The ability to draw accurate conclusions should not require knowledge acquired outside of classroom instruction.

√ **The attached information sheet *must* be completed for scoring purposes. It is also useful as a "check" to evaluate the appropriateness of the selected topic and experiments for this task.**

Teacher _____ **School** _____

Information for Scoring
Problem Solving

This information is necessary for scoring student responses and should be completed before the assessment is administered. Answer each question as completely as possible.

☞ *Science Content Focus (be specific):*

☞ *What responsibilities does the student have during the task administration? (For example, do students observe while the teacher demonstrates the experiment, or are students responsible for conducting the experiments? Are students required to "retell" or "restate" the Problem, the Procedures, and the Results, or are students required to define and formulate each of the scored outcomes without teacher demonstration or explanation?)*

Problem Solving: *Information for Scoring*
 Project funded by The National Science Foundation.

☞ *What portion of the task, if any, is collaborative?*

☞ *Describe the procedures you followed in introducing the experiment and task to students.*

☞ *What scientific concepts are being demonstrated?*

☞ *What conclusions about the scientific concepts demonstrated in the experiment do you expect students to draw?*

☞ *What general information/strategies do you expect students to consider in a plan for investigation?*

Student Worksheet

PROBLEM SOLVING

ASSESSMENT

Problem Solving
Student Response Worksheet

Student Name: _____

School Name: _____

System Name: _____

Teacher Name: _____

Grade Level: _____ Date: _____

After the experiment has been introduced, answer the questions on the following pages.

REMEMBER:

1. Answer each question as carefully and clearly as possible.

2. Your answers will demonstrate
   - your ability to state a hypothesis or research question
   - your ability to explain the procedures or steps that will answer the question/problem
   - your ability to report and explain the results of the experiment

Title of Investigation: _____

Identifies the Problem/Hypothesis

1. Explain the problem or state the research question/hypothesis. What must you find out? What is your hypothesis?

Plans/Reports the Procedures

2. Explain the procedures used to find an answer to the research question or to solve the problem. Be sure to explain how the data were collected.

Problem Solving: *Student Response Worksheet*

Analyzes the Results/Draws Conclusions

3. Describe the results of the experiment. What did you observe? What happened?

4. What scientific information did the results of the experiment demonstrate? What did you learn from this investigation?

5. Do the results agree with the hypothesis? Were you able to find out what you wanted to know? Explain.

If not, explain why.

In contrast to *Letter Writing* or *Science Observation*, an analytic scoring guide or rubric for *Problem Solving* was developed by the school team. They felt strongly that there were three important pieces of summary information needed by any of the stakeholders identified in the model or by any individuals interested in how students implemented scientific methods in answering real-world problems. The scoring guide for this assessment (see next page) describes the three performance areas: Identifying the problem or framing an hypothesis, planning and reporting procedures, and analyzing or drawing conclusions. Each of these performance categories is further described with three levels of performance. The reason that there are only three performance levels within each category reflects the inability of expert judges to sort work within category into more levels while still retaining agreement across judges.

Thus, rather than having overall scores describing overall or general performance in a holistic sense, this assessment activity yields specific and focused pieces of information that can then be combined if desired to provide a general or overall summary performance level.

Examples of Scored Student Work

The scoring guide for *Problem Solving* is analytic with three distinct foci: Identifies the Problem/Hypothesis (2, 1, 0), Plans/Reports Procedures (2,1,0), and Analyzes the Results/Draws Conclusions (2,1,0). It may be helpful to refer to page 96 for further discussion of these score points before reviewing the exemplars that follow. The exemplars are presented by analytic element rather than on total composite score.

Scoring Guide Problem Solving	
	Identifies the Problem/Hypothesis
2	The student defines, formulates, or explains the problem or task and /or poses a hypothesis or research question. The statement is clear, concise, and reasonable, given the stimulus/stipulations.
1	The student defines, formulates, or explains the problem or task and /or poses a hypothesis or research question, but the hypothesis is only partially developed or the problem is partially misunderstood. The response may be too general or narrow, given the stimulus.
0	The student misunderstands the task and does not state a reasonable hypothesis, given the stimulus/stipulations, and/or does not define the problem.
	Plans/Reports Procedures
2	The student defines, formulates, or explains the procedures that have been or will be followed. The approach to the problem is organized, and procedures are explained/ documented; data collection strategies are appropriate for the experiment/activity. The explanation is accurate and complete and could lead to an answer to the problem if implemented correctly.
1	The student defines, formulates, or explains the procedures that have been or will be followed, but the procedures are partially incorrect or incomplete. The approach to the problem may be unevenly organized and procedures and/or data collection strategies show some errors in logic and completeness. There is evidence that part of the plan addresses at least part of the problem.
0	The student misunderstands the task and there is not a reasonable plan of action, given the hypothesis or given the stimulus/stipulations. The plan would not lead to an answer to the problem/hypothesis if implemented correctly; there are no data collection strategies reported. The student does not offer procedures.
	Analyzes the Results/Draws Conclusions
2	The student reaches a logical result/conclusion that relates to the original problem/hypothesis. There is evidence that the appropriate data needed to solve the problem/hypothesis were used. The explanation of the results of the experiment/activity is clear and organized.
1	The student reaches a result/conclusion that relates, in part, to the original problem/hypothesis. There may be evidence that some inappropriate or incomplete data were used to solve the problem/hypothesis. The explanation of the results may be unorganized.
0	The student reaches an illogical result/conclusion or one that does not relate to the original problem/hypothesis. There is evidence that the student misunderstands the task. The student does not offer a result/conclusion.

Examples: Problems/Hypothesis

Student G0122 completed this assessment after a two-week unit on plants also described on page X. The concepts and vocabulary included in this unit were that plants made their own food using photosynthesis, a process in which light energy, carbon dioxide, and water change into food and oxygen in a chloroplast cell. The chloroplast cell has chlorophyll, which makes the leaves of the plant green. To use this food, the plant uses respiration, in which food and oxygen are changed into carbon dioxide, energy, and water. The plant receives sunlight through its leaves. Water and minerals are transported from the roots to the leaves through tubes called xylem. Oxygen and carbon dioxide enter the leaves through tiny holes called stomata. Two experiments were conducted. First, the students attempted to find out what would happen if the stomata of a plant were blocked, or if air was limited. This activity was done in groups of 5 or 6. Each group had three bedding plants. One was the control plant. One plant was coated with petroleum jelly. One plant was enclosed in a plastic bag. The plants were observed for one week. In the second experiment, students were to find the xylem and observe them transporting liquid. Each group had celery and a cup of water colored with red food coloring. The students were to cut off a piece of the celery from the bottom of the stalk and insert it into the colored water. After waiting about 15 minutes, they took the celery out of the water and cut the celery to see if they could find the xylem.)

Student G0122's work earned a score of 2 because it explains an hypothesis, it is clear, concise, and reasonable.

(A) Question: What do you want to find out?

I want to find out how long it takes the food coloring to go up the celery, and see how red it gets.

(B) Hypothesis: What do you think you will find out?

I think the xylem tubes will take the food coloring all the way up to the leaves.

(C) Procedure: List steps you will use to find out.

I will watch the celery, and keep track of the time to see how long it takes.

(D) Results: What happened in your first attempt?

The celery didn't turn very red. It just got a little pink.

(E) Analyze Results: Did you try a different approach? Who or why not?

Yes, we did because we wanted to find out more things.

If you tried a different approach, what did you do differently in your second attempt?

We tried taking just one xylem in the water, and 5 xylems stuck together.

(F) Conclusions: What did you learn and how can you use this information?

I learned that the celery takes in water very quickly. And if you were trying to keep the celery wet you'd have to change the water alot.

(G) Share your findings with the group.

I found out that when we tried to see if the leaves got read by putting them in water, it cut off the oxygen, and the leaves turned black.

Student C0079 completed this assessment after finishing a one-hour lesson on blood types. During this lesson, two experiments were completed. The first activity entailed students working in groups of 4 to determine which blood types are compatible with others and which are not. First, the students were given four cups: (1) water and red food coloring designated as blood type A; (2) one with water and blue food coloring labeled as blood type B, (3) one with water, blue, and red food coloring labeled as blood type AB, and (4) one with plain water labeled as blood type O. They then combined a drop of each of the liquids with a drop from each of the other cups and charted the results. From this experiment, they were to identify the "universal donor."

Student C0079's response was given a score of 1. It included a partially developed hypothesis that indicated that the content was partially misunderstood.

(A) Question: What do you want to find out?

What blood type can give to all other blood types. What blood type can receive from all others

(B) Hypothesis: What do you think you will find out?

I think this is what I am going to found out what kind of blood would you need what kind of blood if you go to the hospital.

(C) Procedure: List steps you will use to find out.

You need 3 cups of food cloring and cup of water, then you some eye drips. You mix all the stuf together with the eye dropper.

(D) Results: What happened in your first attempt?

The color was darker and did not get the right results.

(E) Analyze Results: Did you try a different approach? Who or why not?

Ms. Hackett told us to but two when we go with Trell cause colo is darker then everybody else.

If you tried a different approach, what did you do differently in your second attempt?

Put two more drops color water.

(F) Conclusions: What did you learn and how can you use this information?

If I was a person to take care of blood I would know where I put the blood in

(G) Share your findings with the group.

	A	B	AB	OA
A	A	AB	AB	A
B	AB	B	AB	B
AB	A	AB	ABAB	AB
O	A	B	AB	O

Student G0105 reported on the same experiment as Student G0122. Student G0105 earned a score of 0 because the response did not state a reasonable hypothesis.

(A) Question: What do you want to find out?

What happens?

(B) Hypothesis: What do you think you will find out?

not much

(C) Procedure: List steps you will use to find out.

cut off the bottom of the celery and put in in a cap of water

(D) Results: What happened in your first attempt?

(no response)

(E) Analyze Results: Did you try a different approach? Who or why not?

We stuck some veins in the water and work better because the veins wernt clog

If you tried a different approach, what did you do differently in your second attempt?

stuck the top of the plant in the water and it also sucked up water

(F) Conclusions: What did you learn and how can you use this information?

That the top of the plant can suck up water

(G) Share your findings with the group.

We found out that when you take off the skin the zylems work better

Examples: Plans/Reports Procedures

Student C0059 also reported on the one-hour blood type lesson described on page 98. This response earned a score of 2 because it explains the procedure to be followed—organized plan—and shows data collection.

(A) Question: What do you want to find out?

What blood type can give all other blood types? What blood type can recive from all others?

(B) Hypothesis: What do you think you will find out?

I think 0 type can give to all other and AB can recive all others.

(C) Procedure: List steps you will use to find out.

First me and my partner are going to get a piece of paper and four droppers and cups. Then we are going to fill cups of differ color food coloring. Then we will mix say red wich would be A and Blue wich be B types and find out the results.

(D) Results: What happened in your first attempt?

In our first try A blood was all you could see even when we mixed it.

(E) Analyze Results: Did you try a different approach? Who or why not?

Yes, because A type was all you could see.

If you tried a different approach, what did you do differently in your second attempt?

In our second attempt we put more of each color and it started being other types besides A.

(F) Conclusions: What did you learn and how can you use this information?

I learned that O types can go into any other type and AB type can take any other. I could use it if I every go to a doctor school or a test.

(G) Share your findings with the group.

I found out that when we tried to see if the leaves got read by putting them in water, it cut off the oxygen, and the leaves turned black.

Student G0056 reported on two experiments on the topic of gravity. The students were to draw conclusions that helped them to define the law of gravity. They conducted experiments in which they compared the gravitational force of objects of different sizes, shapes, and masses, dropped from the same height and their rate of descent. They also conducted experiments in which they used items of equal size, shape and mass, dropping them from differing heights and noting the rate of descent. Then the students drew some conclusions incorporating their findings as they related to the law of gravity.

This student evidence was given a score of 1 because it was incomplete. It contained no mention of how data was collected. It was clearly imprecise.

(A) Question: What do you want to find out?

Does size or weight affect how Fast something will Fall?

(B) Hypothesis: What do you think you will find out?

I predict that iF it is heavier than the other it will Fall Faster than the other thing will.

(C) Procedure: List steps you will use to find out.

Choose 5 obJects of different weights and sises. Drop obJects from same place, see how fast thay land

(D) Results: What happened in your first attempt?

The objects Fall at the same time

(E) Analyze Results: Did you try a different approach? Who or why not?

Yes, because the teacher said to.

If you tried a different approach, what did you do differently in your second attempt?

Nock them off the desk.

(F) Conclusions: What did you learn and how can you use this information?

If is was a train lady

(G) Share your findings with the group.

Student R0074 reported on a two-week unit devoted to the study of different types of drugs and the kinds of effects they can have on the human body. The focus of this experiment was to determine what effects caffeine would have on heart rate, student performance, and behavior. During the first week, the students recorded their pulse count at designated times during the day. They also kept a record of the number of correct answers on 5 samples each of mental mathematics, listening, and dictation activities. During the second week, the students received 6 ounces of cola containing caffeine at both 9:00 a.m. and 12:30 p.m. Pulse counts were again recorded at just prior to receiving the caffeine, 30 minutes after and 90 minutes after). Records were also kept of the number of correct answers on the mental mathematics, listening, and dictation activities.

This evidence was scored 0 because it did not present a plan as requested.

(A) Question: What do you want to find out?

We wamt to find out how caffeine affects stduent performance.

(B) Hypothesis: What do you think you will find out?

Some will do better performance. Some will do a very poor performances.

(C) Procedure: List steps you will use to find out.

How much difference in the weeks

(D) Results: What happened in your first attempt?

(no response)

(E) Analyze Results: Did you try a different approach? Who or why not?

No because we just begane.

If you tried a different approach, what did you do differently in your second attempt?

The same thing but have the color more than two times.

(F) Conclusions: What did you learn and how can you use this information?

I learned that id does not have a effect on me and my work.

(G) Share your findings with the group.

(No response)

Examples: Analyzes the Results/Draws Conclusions

Student G0059 also reported on the lesson on blood types described on page 98. This student response was given a score of 2. It contains a logical conclusion addressing original hypothesis. Imprevise language was used but it was understandable. And, appropriate data was used.

(A) Question: What do you want to find out?

What blood type can give all other blood types? What blood type can recive from all others?

(B) Hypothesis: What do you think you will find out?

I think 0 type can give to all other and AB can recive all others.

(C) Procedure: List steps you will use to find out.

First me and my partner are going to get a piece of paper and four droppers and cups. Then we are going to fill cups of differ color food coloring. Then we will mix say red wich would be A and Blue wich be B types and find out the results.

(D) Results: What happened in your first attempt?

In our first try A blood was all you could see even when we mixed it.

(E) Analyze Results: Did you try a different approach? Who or why not?

Yes, because A type was all you could see.

If you tried a different approach, what did you do differently in your second attempt?

In our second attempt we put more of each color and it started being other types besides A.

(F) Conclusions: What did you learn and how can you use this information?

I learned that O types can go into any other type and AB type can take any other. I could use it if I every go to a doctor school or a test.

(G) Share your findings with the group.

I found out that when we tried to see if the leaves got read by putting them in water, it cut off the oxygen, and the leaves turned black.

Student M0010 reported on an instructional activity that lasted two weeks. The focus of the unit was design and its effect on whether and for how long a paper airplane will stay aloft. This response earned a score of 1 because it relates, in part, to the original hypothesis but it is not as well organized as would earn a 2.

(A) Question: What do you want to find out?

Design a paper airplan that will fly father than anyone else.

(B) Hypothesis: What do you think you will find out?

How far can it fly.

(C) Procedure: List steps you will use to find out.

We made a air plane. We test fly it. We change are design.

(D) Results: What happened in your first attempt?

We both had a airplane go around.

(E) Analyze Results: Did you try a different approach? Who or why not?

Yes: Becaus we didn't go far. We change the design.

If you tried a different approach, what did you do differently in your second attempt?

We put gem clip on are paper airplane.

(F) Conclusions: What did you learn and how can you use this information?

We yseh gemclip to make it go father. in

(G) Share your findings with the group.

(No response)

Student C0077 reported on the same experiment as students C0079 and C0059. This response earned a score of 0 because it contained errors in logic and completeness. It does, however, address part of the problem.

(A) Question: What do you want to find out?

A what 'blood' types can give to all other blood types? O blood B what blood tuypes can recives from all over? Ab blood

(B) Hypothesis: What do you think you will find out?

I think I will find out what do they do when you mix each color together

(C) Procedure: List steps you will use to find out.

I will mix the colors together on a pice of paper. the colors I will use are purplse, clue, red 3 water frist the colors are in cups reds in one then blue is in one then purple in one and water in one together with a eye drop

(D) Results: What happened in your first attempt?

On the first try it went well

(E) Analyze Results: Did you try a different approach? Who or why not?

We put to muuch when we were mixing with the clear.

If you tried a different approach, what did you do differently in your second attempt?

We did not put much on the paper

(F) Conclusions: What did you learn and how can you use this information?

I learned that you had to really mix them up to get another color. It is hard to mix blood. When the have to mix it but AB blood are not hard to mix.

(G) Share your findings with the group.

(no response)

8

Comparison of Experiments

This chapter includes the *Comparison of Experiments* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form and are available for reprint and use.

Development History

The *Comparison of Experiments* assessment entry for the assessment portfolio is an extension of the Problem Solving task in that it creates a context in which two related but different experiments are combined to produce insights and understanding in a more general and broad-based arena. This task was developed by the Marietta school team in an attempt to challenge students at the meta level of science thinking. The intent was to have the students conduct two experiments pertaining to specific science content and then to draw conclusions or to generalize about that content in ways only likely because of science understanding stimulated through differences and similarities in related scientific endeavors.

Relative to the dimensions of time, content complexity, stimulus complexity, and response complexity, it scores as follows:

Attribute	Score
Instructional Time	Four to eight weeks (multiple units/lessons)
Content Complexity	Complex
Stimulus Complexity	Complex
Response Complexity	Complex

Compared with the dimensions of time, content complexity, stimulus complexity, and response complexity present in previous assessment entries, *Comparison of Experiments* is complex with respect to content complexity, stimulus complexity, and response complexity. Further, it is the most demanding of all the assessment entries relative to requisite instructional time preceding the assessment itself.

Comparison of Experiments addressed the project goals of "reflective thinkers and self-evaluators," "creative and strategic thinkers," "self-directed learners," "effective communicators," "experiential learners," and "responsible global citizens." The structure of the task is essentially that a student conduct two experiments selected by the teacher to elicit certain generalizations in support of increased sophistication in science and that each student use the information gleaned from each of the separate experiments to deduce generalizations of merit that are only likely to have come from the comparative process.

Like *Problem Solving*, *Comparison of Experiments* is clearly a science assessment activity. It is not intended to be used in mathematics although mathematics knowledge and skills are certainly requisite for quality work.

Teacher Instructions*(Second Revision)*

COMPARISON OF EXPERIMENTS ASSESSMENT

Teacher Instructions: Comparison of Experiments

Overview

Students will observe the results of two experiments pertaining to specific science content (topic) and draw conclusions about the scientific principles at work. Additionally, students will be asked to note the important similarities and differences between the experiments and the possible applications of this information in settings outside the school.

Purpose

The task was designed to evaluate students'

1) Understanding of the scientific concepts presented.
2) Ability to draw conclusions about the concepts presented, based on the results of the two experiments.
3) Understanding of how the information "fits" into our society at large.
4) Ability to communicate this information effectively in written form.

Planning

√ **To plan the science context (experiments) for this evaluation, the following must be considered:**

1) The experiments must demonstrate scientific concepts appropriate for the science curriculum. It is important to consider what the two experiments demonstrate. For example, does experiment #1 manipulate variable x while experiment #2 manipulates variable y, or do both experiments demonstrate a different characteristic of a scientific principle?

√ **Read the rubric before selecting the experiments; the knowledge will help in guiding your selection.**

2) The two experiments must allow for meaningful comparisons of similarities and differences.

3) The scientific concepts demonstrated in the experiments must allow for accurate conclusions. The ability to draw accurate conclusions should not require knowledge acquired outside of classroom instruction.

4) The scientific principles demonstrated must have a practical application. Discussions concerning "real life" application and possible interested audiences should be included during instruction.

√ **The attached information sheet *must* be completed for scoring purposes. It is also useful as a "check" to evaluate the appropriateness of the selected topic and experiments for this task.**

Teacher _____ **School** _____

Information for Scoring
Comparison of Experiments

This information is necessary for scoring student responses and should be completed before the assessment is administered. Answer each question as completely as possible.

☞ *Science Content Focus (be specific):*

☞ *What responsibilities does the student have during the task administration? (Do students observe while the teacher demonstrates the experiments or are students responsible for conducting the experiments?)*

☞ *What portion of the task, if any, is collaborative?*

Comparison of Experiments: *Information for Scoring*
 Project funded by The National Science Foundation.

☞ *Briefly address the following questions: What is the purpose of this assessment as you have designed the experiments? How does this assessment "fit" with the curriculum? (For example, the experiments may be novel while the concepts demonstrated are part of the instructional sequence. Hence, the teacher may use this task to determine whether or not the student can transfer learning to novel situations -- to explain what is observed -- either as a formative evaluation or as a summative evaluation).*

☞ *What science concepts are being demonstrated?*

☞ *What are the pertinent similarities between the experiments that students might observe?*

☞ *What are the pertinent differences between the experiments that students might observe?*

☞ *What conclusions about the scientific concepts demonstrated in the experiments would you expect students to draw?*

☞ *During instruction, what "real life" applications and suggestions of appropriate audiences for the information were presented?*

☞ *What science terms would you expect students to use in their responses to the questions on the student worksheet?*

☞ *Describe 1) the steps, and 2) the results of Experiment #1 and Experiment #2. (If additional space is needed, please use another sheet of paper and attach it to the completed* ***Information for Scoring Form****.)*

Comparison Of Experiments
Student Response Worksheet

Student Name: _____

School Name: _____

System Name: _____

Teacher Name: _____

Grade Level: _____ Date: _____

After both experiments have been completed, answer the five questions on the following pages.

REMEMBER:

1. Write in complete sentences.
2. Use science vocabulary.
3. Read each question carefully.

1. What were some important similarities in both experiments?

2. What were some important differences between the two experiments?

3. What scientific conclusions can you make about the topic? To answer this question you should

   - *Briefly* report the results of the experiments (using only one or two sentences).
   - Then discuss the scientific principle(s) that would explain the results.

4. How is this topic connected to the world outside of school?

5. The results of the experiments and the conclusions you drew would be of interest to people in what job or profession. Explain *why* they would be interested.

The scoring guide for *Comparison of Experiments* is almost a combined holistic and analytic model (see next page). The three areas of focus (analytic elements) are Understands Scientific Concepts, Extends Learning, and Communicates Understanding. Within each of these areas of focus holistic judgements are defined by multiple elements. The score range is from 1 to 4 for each area of focus with a zero score when there is not even enough evidence to award a score of 1 (or minimal). In this case, that zero score would also accommodate "off task" performances.

Scoring Guide: Comparison of Experiments			
Optimal: Score 4	**Good:** Score 3	**Satisfactory:** Score 2	**Minimal:** Score 1
Understands Scientific Concepts: *Questions 1, 2, and 3*			
• identifies similarities pertinent to concepts being demonstrated • identifies differences pertinent to concepts being demonstrated • draws accurate conclusions about the scientific concepts demonstrated in the experiments	• identifies similarities pertinent to concepts being demonstrated **OR** identifies differences pertinent to concepts being demonstrated • draws conclusions about the scientific concepts demonstrated in the experiments with minor errors	• identifies similarities pertinent to concepts being demonstrated **OR** identifies differences pertinent to concepts being demonstrated **OR** draws conclusions about the scientific concepts demonstrated in the experiments with minor errors *(gives more general response; may report results but not concept)*	• identifies similarities **OR** identifies differences **OR** draws conclusions about the scientific concepts demonstrated in the experiments with minor errors • evidences misunderstanding of concepts
Extends Learning: *Questions 4 and 5*			
• offers appropriate application to "real life" • suggests appropriate audience	• offers appropriate application to "real life" • suggests appropriate audience (may be appropriate but general)	• offers appropriate application to "real life" **OR** suggests appropriate audience	• is not able to appropriately apply to "real life" or to additional audience but does write response
Communicates Understanding: *Questions 1, 2, 3, and 4*			
• explanations are clearly stated and complete • explanations include science terms	• explanations are clear but may be unevenly developed • explanations include science terms	• explanations are clear but may be unevenly developed • explanations include science terms	• explanations may not be easily understood • science terms may not be used or may be used incorrectly
A score of "0" should be given if the standards for a minimal score of "1" are not met.			

Examples of Scored Student Work

This assessment is perhaps the most complex of any developed for this portfolio project because not only does the content vary from classroom to classroom but it requires two activities or experiments representing some kind of teacher-identified meaningful contrast. The samples of student work on *Comparison of Experiments* included in the following pages are intended to provide the essence of the difference in levels of quality rather than to provide definitive descriptions of each possible score point. These samples provide the essence of the difference in levels of quality.

The three elements on the scoring guide for Problem Solving are: *Understands Scientific Concepts*, *Extends Learning*, and *Communicates Understanding*. The range of performance was scored from 4 (Optimal) to 1 (Minimal). You may want to turn to page 120 to review the scoring rubric for Comparison of Experiments.

Each of the examples of student work presented here came from the same classroom. The unit of study summarized through this assessment was one on classifying invertebrates and vertebrates. The unit lasted five weeks. Two experiments were conducted. The first experiment was to determine the effectiveness and purpose of downy feathers and large wing feathers of birds. The students were instructed to stand on top of a chair and drop each kind of feather to observe how it fell. The students were to discuss which feather would be better for the body or a bird and which feather would be better for the bird's wing. The second experiment required that the students compare chicken and beef bones and to relate these differences to the necessities of the lives of these animals. The *Comparison of Experiments* assessment was given one day after the second experiment had been completed to give the students time for reflection.

Example 1

Student G0119 received a score of "4" for Understands Scientific Concepts, a "2" for Extends Learning, and a "4" for Communicates Understanding.

1. What were some similarities of both experiments?

   They both were dealing with the sturte of birds. Or parts of the birds body,

2. What were some differences between the two experiments?

   One was to see how they keep themselves warm and which feather would flot to the ground the fastest. The other was talking about how there bone were made.

3. Having completed both experiments, what conclusions about the topic can you make now?

   That there bones are made to suit there lifestyle. Like they are able tpo move the right way to the way they fly. I also thought about having the feathers. I never knew there were two kinds of feathers. But it make since. What good would the warmth feathers do on the outside?

4. Can you think of another experiment for this topic?

   Yes, how they eat. Do they play with there good like whales or just eat it in one bit like frogs.

5. What would you use?

   A pen and paper, binoculars, and a camera.

6. How would you do this?

   Find a nest and a good spot were I can seem and sit and wait to they start eating. I would record what I saw take pictures and study them everyday.

7. If you were given your materials, would you try this experiment on your own?

   Yes, I would.

8. Why or why not?

   Because I think it would be neat to find out the way they eat. It would exciting almost.

9. How is this topic connected to the world outside of school?

   Because there are birds in the outside world and there are people who study them.

10. Who do you think might be interested in the data and conclusions you drew on this topic?

    Mrs. Crawford, scients (sic), *myself, and other teachers.*

Example 2

Student G0102 received a score of "4" for Understands Scientific Concepts, a low "2" for Extends Learning, and a "3 ' for Communicates Understanding.

1. What were some similarities of both experiments?

   The objects that we used both came from animals and both objects could break. Both objects were bones, the feather had a bone in it.

2. What were some differences between the two experiments?

   The objects did not come from the same animals. One of the bones could not break, that was the beef bone. The feather bone was hollow the chicken bone and beef bone had bone maroe in them.

3. Having completed both experiments, what conclusions about the topic can you make now?

   I learned that bird bones are hollow and some aren't. Mammels don't have hollow bones. The reason why is because mamels need to be strong so the can run and other stuff that involves standing up. I also learned that most birds have two kinds of feathers, small ones to keep them warm and big ones to help them fly.

4. Can you think of another experiment for this topic?

   Yes, and I would like to go into a zoo or a forest, find different types of animals and there bones and see what the differences are. See if they're hollow or not.

5. What would you use?

   I would use a cage, gloves, gun just in case something attackes me and I would have two or three helpers.

6. How would you do this?

   I would trap an animal, put it in a cage and examine it. I would also look around for bones from animals that had already died.

7. If you were given your materials, would you try this experiment on your own?

   Yes, I would when I was older because I don't know how to do all that stuff at this age.

8. Why or why not?

   Because I want to learn about different kinds of animals and what there bones are good for. Like if they fly they might need hollow bones, if ther mammels they wouldn't need hollow bones.

9. How is this topic connected to the world outside of school?

   Birds and mammals really do live outside of school. They also get killed.

10. Who do you think might be interested in the data and conclusions you drew on this topic?

    A forest ranger or a person who really works hard to keep the forest clean.

Example 3

Student G0121 received a score of "2" for Understands Scientific Concepts, a "1" for Extends Learning, and a high "1" for Communicates Understanding.

1. What were some similarities of both experiments?

   both of them were about how they are made and how them work

2. What were some differences between the two experiments?

   one used bones and the other one used feathers

3. Having completed both experiments, what conclusions about the topic can you make now?

   Birds bones are softer than cow bones. Big feathers helf bird fly better than small feathers.

4. Can you think of another experiment for this topic?

   How can a mouse eat a hole in a wall.

5. What would you use?

   A cage a borad and a mouse.

6. How would you do this?

   Have a wood borad in a cage with a mouse and see if it would eat it.

7. If you were given your materials, would you try this experiment on your own?

   Yes, I would.

8. Why or why not?

   Yes Because it seems like a lot of fun but my mom will not let me do this at home so I would have to do this at school.

9. How is this topic connected to the world outside of school?

   It is not ouside of school im going to do this in school.

10. Who do you think might be interested in the data and conclusions you drew on this topic?

    My class mates and brothers would and of chorse me.

9

Continuum of Progress

This chapter includes the *Continuum of Progress* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form and readers are invited to reprint and use the materials.

Development History

The fifth assessment entry developed for the assessment portfolio developed was called the *Continuum of Progress Towards Goals* assessment, or simply *Continuum of Progress*. This task is designed for either science or mathematics content. This task requires that the student be reflective about his or her growth in understanding as well as how the new content materials might be useful in further study and in life outside school. This task looks and acts very much like a survey. An important distinction between typical surveys and the *Continuum of Progress* is that there is a scoring rubric for the student responses. Therefore, the responses are judged for quality rather than simply coded as descriptors as in a traditional survey. It also focuses on instruction areas of interest such as prior knowledge from the student's perspective, judgments about what the student learned within a defined time frame, student goals for learning, student's plans for continued learning, and so forth. *Continuum of Progress* is intended for use at the end of a major block of instructional time (quarter, semester, year). Its focus is on what, from among the instructional experiences available during a major period of time, each individual student particularly valued, benefitted from, and intends to expand upon.

Relative to the dimensions of time, content complexity, stimulus complexity, and response complexity, it scores as follows:

Attribute	Score
Instructional Time	Multiple units over large periods of time (quarter, semester, year)
Content Complexity	Complex and Variable
Stimulus Complexity	Complex and Variable
Response Complexity	Ranges from simple to complex over the questions in the assessment

Compared with the dimensions of time, content complexity, stimulus complexity, and response complexity present in *Problem Solving, Continuum of Progress* is less focused and is unlikely to give comparable information on a given unit of instruction across students because it is intended to elicit from students evidence about their particular and individualistic values and perceived benefits over a large pool of learning opportunities. However, the judgements would be comparable.

Continuum of Progress was developed by the Dade County school team. This assessment strategy addressed all of the project goals of "effective communicators," "responsible global citizens," and "reflective thinkers and self-evaluators." The mode of communication is writing.

Teacher Instructions

CONTINUUM OF PROGRESS TOWARDS GOALS ASSESSMENT

Teacher Instructions: Continuum of Progress Toward Goals

Overview

The students are introduced to a topic of study in science or math. After the topic has been introduced and taught for two or three class periods, the students will complete questions 1 through 3 on the assessment survey. These questions evaluate the students' ability to formulate questions appropriate for the topic being studied and plan a strategy for finding the answers to their questions. When the research plan has been implemented, the students complete questions 4 and 5 in the assessment survey. These questions evaluate the students' ability to summarize those findings related to their research questions and areas of interest.

Purpose

The task was designed to evaluate the students'

1) Ability to design focused, appropriate, well-formulated questions related to an area of study
2) Ability to develop multiple, systematic, and appropriate strategies for gathering information that will answer the research questions
3) Ability to accurately summarize and report their findings and demonstrate an understanding of the content researched
4) Ability to apply their findings in different contexts by determining an appropriate audience for the information learned and plausible life applications as well as generating new questions for research.

Planning

√ **These directions should be followed:**

1) After a topic of study has been introduced and taught for two to three days, the students should be introduced to the *Questionnaire/Survey Assessment, Part I*.

   - Review the Part I questions with the students. As you explain question 3, remind students that they should provide multiple strategies and resources for gathering information to answer their questions.
   - Share the rubric with the students, explaining what you will be looking for in their responses to the questions in Part I.
   - Have the students fill in their topic in all of the appropriate blanks (beginning sentence, questions 1a, 1b, 2a).
   - The students should then complete Part I. Allow enough time for the students to respond thoughtfully and completely.

2) Allow students to carry out their research plan as noted in question 3.

3) After students have gathered the necessary information, they should be instructed to continue with the *Questionnaire/Survey Assessment, Part II*.

   - Review the Part II questions with the students.
   - Share the rubric with the students, explaining what you will be looking for in their responses to the questions in Part II.
   - Have the students fill in their topic in the blank at the top of the page.
   - The students should then complete Part II.

√ **The attached information sheet MUST be completed for scoring purposes.**

Teacher _____ **School** _____

Information for Scoring
Continuum of Progress Towards Goals

This information is necessary for scoring student responses and should be completed before the assessment is administered. Answer each question as completely as possible.

☞ *What is the topic of study to be introduced?*

☞ *What information, such as content (concepts and skills) and vocabulary will be introduced prior to the administration of the task? the experiments?)*

☞ *What is the length of instructional time prior to the assessment?* _____

Continuum of Progress Toward Goals: *Information for Scoring*

CONTINUUM OF PROGRESS TOWARD GOALS

Student Name: _____

School Name: _____

System Name: _____

Teacher Name: _____

Grade: _____ Date: _____

Continuum of Progress Toward Goals
Student Response Worksheet

PART I

We are beginning/continuing a study of _____.

Think about and answer the following survey questions:

1a) What did you already know about _____ before this school year?

_____

_____

_____

_____

_____

1b) What did you learn in class this year about _____?

_____

_____

_____

_____

_____

2a) What (more) would you like to know about _____?

_____

_____

_____

_____

_____

2b) What are some questions you could ask to investigate what you would like to know about the topic?

2c) How are your new questions related to what you already know?

3) What are some ways you might find answers to your question(s)? Design a strategy to use to answer your question(s). Write out your plan.

PART II

You have now completed your study of _____.

Answer the following survey questions:

4a) What did you find out that you wanted to know? What did you learn about the topic that answered your questions? (See question 2b.)

4b) What other things did you learn that you did not know before?

5a) What new questions were raised as a result of your study?

5b) How is the information you studied useful to you? How can you apply some of what you learned to what you do every day?

5c) Who might be able to use the information you learned from your study? Why?

Continuum of Progress Towards Goals: *Student Response Worksheet*

The scoring guide for *Continuum of Progress Towards Goals* assessment is analytical. The areas of focus are: Develops Focus, Develops Strategies, Summarizes Findings, and Applies Findings. These, as areas of focus model the scientific and mathematical habits of mind even though they are not, in this assessment, specifically focused on discrete content areas. A student receives a check mark for each element within a focus area as appropriate. Thus, the range of points varies for each area of focus depending upon the elements listed (4 to 7). The total number of check marks across all four elements of focus is 0 to 20. No "off task" score is provided within this scoring schema.

Scoring Guide: Continuum of Progress Towards Goals

	PART I
	Develops Focus ***(Read student's responses to questions 1 and 2.)***
	Question(s) or statement(s) indicate what student wants to find out.
	Question(s) or statement(s) are appropriate for the stated topic/content.
	The scope of the question(s) or statement(s) is appropriately limited.
	The question or statement is well formulated OR most of the question(s) or statement(s) are well formulated.
	Develops Strategies ***(Read student's response to question 3.)***
	There is a plan to gather information.
	The plan is detailed and systematic.
	The plan is appropriate given content, resources, and time constraints.
	The plan is clearly linked to the question(s)/statement(s) posed.
	There is evidence of multiple strategies.
	PART II
	Summarizes Findings ***(Read student's responses to questions 4.)***
	Findings are reported; there is evidence that the student gathered information.
	Findings presented answer/address the question(s)/statement(s) posed (either initial or restated).
	Answers for all of the question(s)/statement(s) have been reported/explained.
	Findings are detailed and well explained.
	Information reported is accurate.
	Content vocabulary is used appropriately.
	Content vocabulary is used extensively.
	Applies Findings ***(Read student's response to question 5.)***
	Questions derived from the findings are reported.
	The new questions are reasonable spin-offs from the information studied.
	Student offers a plausible application for the information studied; information must be accurate.
	Student identifies an appropriate additional audience with an adequate explanation as to why the audience was named.

Continuum of Progress Towards Goals: *Scoring Guide*
 Project funded by the National Science Foundation.

Examples of Scored Student Records

The scoring guide for *Continuum of Progress Towards Goals* assessment is analytical. The areas of focus are: Develops Focus, Develops Strategies, Summarizes Findings, Applies Findings. The range of points varies for each area of focus depending upon the elements listed (4 to 7). The total number of check marks across all four elements of focus is 0 to 20.

Example 1

Student C0076 earned a total score of 15 out of a possible 20. The category scores are 4 for Focus, 5 for Develops Strategy, 5 for Summarizes Findings, and 1 for Applies Findings.

1) What do you already know about *the heart and lungs? I know that the heart is about the size of our fish. I know that the heart pumps blood to all the cells in your body. I know that the liungs can turn black if you smoke alot. Your lungs is in your respiratory system.*

2) What (more) would you like to find out about *cancer -- what causes cancer? Have they found a cure for cancer yet?*

3) What do you think you could find out? Describe your plan in writing. *I will visit the hospital that Athens Regional Hospital Lab. In class I will check out books from our libray and take notes. The heart is part of the circulatory. The heart pushes blood through your heart 24 hours without rest. The heart is the pump of the circulatory system. The heart is a hollow muscle. The bottom chamber of the heart punps blood to the lungs.*

1) Did you find out what you wanted to know? PROVE IT. *Yes I did found that there is know cure to cancer. Lookg on the back. I found out what I need to I know there is not a cure for cancer yet. I know what causes cancer when someting goes wrong with the tiny cells. I got my info from books and I wrote a letter to the American Cancer Society.*

2) How can you apply your findings to "the real world"? *If I was going to be on a oncologists, I can some of the studies of the cancerns in the human body.*

3) You've shared your results with your group. Would you change what you did according to suggestions from the group? What would you change and how would you change it? *I would choose a different topic like aids.*

Example 2

Student F0117 also earned a score of 15 out of 20. Student F0117 earned 4 for Develops Focus, 5 for Develops Strategy, 5 for Summarizes Findings, and 1 for Applies Findings.

1) What do you already know about *the systems of the body. I know that your pancreas are an organ in your body tat produces juices that break down your food to go through the needed processes. Your lungs is in your respiratory system.*

2) What (more) would you like to find out about *the systems of the body -- I would like to know how can you help your body fight off a cold?*

3) What do you think you could find out? Describe your plan in writing. *I think I could inf out by: Going to the libriary and doing research asking my doctor, and researching iun my science book.*

1) Did you find out what you wanted to know? PROVE IT. *By taking vitamins (pills, drinking plenty of liquids, keeping warm and getting alot of rest).*

2) How can you apply your findings to "the real world"? *Next time I get the sniffles I'll know to take a vitamin C pill, drink some guide get covered up in bed and go to sleep.*

3) You've shared your results with your group. Would you change what you did according to suggestions from the group? What would you change and how would you change it? *No, they didn't say I needed to change it.*

Example 3

Student M0007 earned a total composite score of 11 out of 20. The category scores are 3 for Develops Focus, 3 for Develops Strategy, 5 for Summarizes Findings, and 0 for Applies Findings. Thus, this response shows a weakness in transferring or generalizing beyond this particular learning experience.

1) What do you already know about *sun, moon, pants? I know that the Eart is around the Sun. The Sun is bigger.*

2) What (more) would you like to find out about *sun, moon, planets -- sun hot these the Earth*

3) What do you think you could find out? Describe your plan in writing. *I would go to the library and look it up.*

1) Did you find out what you wanted to know? PROVE IT. (no response)

2) How can you apply your findings to "the real world"? *I want to fine out how meane helps can fite on the moon.*

3) You've shared your results with your group. Would you change what you did according to suggestions from the group? What would you change and how would you change it? *No, I well not!*

Example 4

Student D0008 earned a total composite score of 8 out or 20 with a 3 for Develops Focus, 2 for Develops Strategy, 3 for Summarizes Findings, and 0 for Applies Findings.

1) What do you already know about *solar system? I know that the sun is a star. There are 9 planets. Pluto is the coldest plante. Mercury is the botest planet. Jupiter is biggest.*

2) What (more) would you like to find out about *solar system -- I would like to find out how long is the solar system is. How many stare are there. Are there life forms on any other planets. How many rings are around Uranus?*

3) What do you think you could find out? Describe your plan in writing. *I think I could find out if there are any life forms in space by visiting the planets. I could count the rings.*

1) Did you find out what you wanted to know? PROVE IT. *I found out that the moon has 4 phases.*

2) How can you apply your findings to "the real world"? *I can use the moon phases to tell if the day will be longer or shorter.*

3) You've shared your results with your group. Would you change what you did according to suggestions from the group? What would you change and how would you change it? *(no response)*

Example 5

Student G0009 earned a total composite score of 2 out of 20, earning points only in the category of Develops Strategy. This is clearly a weak response.

1) What do you already know about *ecology? air pollution recycling ozone hot*!

2) What (more) would you like to find out about *ecology -- stove we do not know.*

3) What do you think you could find out? Describe your plan in writing. *every every we did a lot of experiments we did researched we diud write reports we cao ecology journal*

1) Did you find out what you wanted to know? PROVE IT. *yes, because I read about it*

2) How can you apply your findings to "the real world"? *yes we can save the enveafy! We study this in the class room*

3) You've shared your results with your group. Would you change what you did according to suggestions from the group? What would you change and how would you change it? *no I would not change it*

10

Retelling

This chapter includes the *Retelling* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form and readers are invited to reprint and use materials.

Development History

This assessment was developed by the project staff rather than a school team. It derived from the instructional strategy in reading. The use of retellings as assessments was not well documented in the literature. However, the MAPS experience in Canada was influential in our decision to try this technique as an assessment tool in mathematics. Essentially, the approach requires that students read or hear a stimulus that has mathematics content in it but that is not explicitly presented as mathematics. The student is then required to translate from narrative text to mathematical language the gist of the stimulus (or story). In the Retelling assessment task, the stimulus is fixed, it is a poem written to parallel Going to St. Ives:[1]

> While on the road to John o'Groats
> I met a couple with seven goats
> Each goat had a ribbon on each of its horns
> And carried on its back a sack of corn
> How many sacks, ribbons, people, and goats
> Were on their way to John o'Groats?

The mathematics underlying this specific and fixed content stimulus was modeling, operations, reasoning, problem solving, and communication.

The students were asked to complete three specific tasks: Write the word problem in their own words (modeling and communication), solve the problem (reasoning, problem solving, operations), and describe how they solved the problem (communication). The project goals addressed across these three products were "reflective thinkers and self-evaluators," "creative and strategic thinkers," "self-directed learners," "effective communicator," and "experiential learner."

Relative to the dimensions of time, content complexity, stimulus complexity, and response complexity, it scores as follows:

Attribute	Score
Instructional Time	Unspecified
Content Complexity	Fixed
Stimulus Complexity	Fixed
Response Complexity	Variable depending upon the student's mathematics sophistication

Compared with the dimensions of time, content complexity, stimulus complexity, and response complexity present in any of the previously described assessments, *Retelling* is more constrained. However, the use of John o'Groats as a stimulus is not required for the task to be useful. In fact, the universe of eligible stimulus materials is infinite. What is critical to the inherent structure of *Retelling* is that translation of some degree of complexity is required. Thus, the use of literature with mathematical content is ideal. The notion here is that the student must make mathematical sense of something that is not clearly "mathematical" in its presentation. The mode of communication is writing. However, drawing or physical models were not prohibited.

[1] A Mother Goose rhyme

In addition to preparing the typical assessment materials (Teacher Instructions, Student Worksheet, Scoring Guide), the project staff who developed *Retelling*, prepared a formal mapping summary relating each element of this task to the project goals. This process of mapping or explicitly linking each task element to the project goals served as a useful reminder of the original purpose for developing the tasks. And, because so often tasks take on a life of their own, the mapping procedure was a good reality check. Whether or not a formal process is necessary is debatable. However, because good portfolio entries are so difficult to construct, it is not unlikely that "nifty" or interesting tasks would likely be included just because they exist and not necessarily because they represent the best measurement practice or the most efficient, useful, and credible source of evidence of student learning. For this reason alone, it is useful to map backwards from the task evidence to be elicited to the reasons for developing the task initially.

Mapping of the Mathematical Retelling Assessment Activity

Write the word problem in your own words in the space below.

Creative and Strategic Thinker

Builds on previous knowledge.
Is able to access information from multiple sources.

Self-directed Learner

Exceeds basic requirements.

Effective Communicator

Can show written evidence of work through narration, description, persuasion, and exposition.

Experiential Learner

Is involved in student-directed activities.
Articulates to audiences.

Solve the problem. Show your work in the space below.

Reflective Thinkers and Self-evaluator

Demonstrates ability to articulate steps (approaches) to problem situation.

Demonstrates ability to recognize the act of transference from one learning situation to another.

Creative and Strategic Thinker

Uses systematic procedures/processes things systematically.

Uses trial and error problem solving.

Has a rational plan.

Builds on previous knowledge.

Demonstrates flexible thinking.

Self-directed Learner

Makes choices and sticks to choices.

Takes initiative.

Tries things in a new way.

Exceeds basic requirements.

Effective Communicator

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 gather information through reading and being read to.

Uses appropriate vocabulary for math and science.

Experiential Learner

Is involved in student-directed activities.

Articulates to audiences.

Effective Collaborator

Responsible Global Citizen

Describe how you solved the problem in your own words in the space below. You may use pictures in addition to words to describe your problem-solving process.

Reflective Thinkers and Self-evaluator

Demonstrates ability to articulate steps (approaches) to problem situation.

Creative and Strategic Thinker

Uses systematic procedures/processes things systematically.
Has a rational plan.
Builds on previous knowledge.
Demonstrates flexible thinking.

Self-directed Learner

Makes choices and sticks to choices.
Takes initiative.
Tries things in a new way.
Exceeds basic requirements.
Uses wait time effectively (finds something meaningful to do after completing tasks).
Moves outside of individual comfort zone.
Tries things in a new way.

Effective Communicator

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 gather information through reading and being read to.
Uses appropriate vocabulary for math and science.

Experiential Learner

Is involved in student-directed activities.
Articulates to audiences.

Effective Collaborator

Responsible Global Citizen

Teacher Instructions

RETELLING APPLIED TO WORD PROBLEMS IN MATHEMATICS ASSESSMENT

John o'Groats Stimulus

While on the road to John o'Groats
I met a couple with seven goats
Each goat had a ribbon on each of its horns
And carried on its back a sack of corn
How many sacks, ribbons, people, and goats
Were on their way to John o'Groats?

Retelling: *Teacher Instructions*
 Project funded by the National Science Foundation.

Retelling Applied To Word Problems In Mathematics
Student Worksheet

Student's Name _____

School _____ Word Problem Category 1 2 3
(attach specific word problem)

Grade _____

Teacher _____ Date _____

Retelling: *Student Response Worksheet*
 Project funded by the National Science Foundation.

STUDENT WORKSHEET

Write the word problem in your own words in the space below.

STUDENT WORKSHEET (continued)

Solve the problem. Show your work in the space below.

STUDENT WORKSHEET (Continued)

Describe how you solved the problem in your own words in the space below. You may use pictures in addition to words to describe your problem-solving process.

Retelling: *Student Response Worksheet*

The scoring guide for *Retelling* is analytic. The focus areas are estimation, retelling the word problem, describing the solution process, and demonstrating mathematical confidence. These areas support the NCTM Standards as well as the project goals. Within each focus area, the possible points range from "0" to "3." Both a total score across all three focus areas and three separate scores may be generated from this assessment.

Points Represent
0 = None 1 = Attempts 2 = Some 3 = Most 4 = Complete

Elements	**Student's Points**
Estimation (if applicable) a. Makes no attempt b. Makes estimate reflecting inappropriate operation(s) c. Makes estimate reflecting appropriate operation(s)	
Retells the word problem a. Identifies problem b. Includes accurate details/facts c. Identifies essential relationships	
Describes the Solution Process a. Identifies needed data b. Identifies an appropriate solution strategy c. Solves problem appropriately d. Checks solution	
Demonstrates Mathematical Confidence a. Uses appropriate mathematical language b. Perseveres in problem-solving attempts	

Retelling: *Scoring Guide*
 Project funded by the National Science Foundation.

Examples of Scored Student Work

This assessment is scored using an analytic guide with four distinct elements: Estimation (3, 2, 1, 0), Retells the Word Problem (3, 2, 1, 0), Describes the Solution Process (3, 2, 1, 0), and Demonstrates Mathematical Confidence (3, 2, 1, 0).

Example 1

Student R0051 earned a score of 4 on each dimension of the analytic for a composite of 16.

Write the word problem in your own words in the space below.

I was on my way to John O'Groats. When I met a couple walking 7 goats. Each goat had a ribbon on each of its horns. And they all were carrys sacks of corn on their backs. How many sacks of corn, ribbons, people and goats were on their way to John O'Groats?

Solve the problem. Show your work in the space below.

7 goats
+ 7 sacks of corn
14
14 bows
28
+ 2 people + 1 were going to O'Groats

(plus picture of "me" plus two people plus seven goats plus seven sacks of corn plus 14 bows)

Describe how you solved the problem in your own words in the space below. You may use pictures in addition to words to describe your problem-solving process.

In the problem it said two people were to O'Groats. So I drew to people. Then they had 7 goats with sacks of corn on backs and bows on their horns, So I added the total amount of corn, people, sacks, goats, and bows and got 31. So, 31 people, sacks, goats, and bows were heading to O'Groats.)

Example 2

Students D0019 earned 3 points on each of the dimensions for a total score of 12.

Write the word problem in your own words in the space below. *Two people were on they way to John O-Groats. With them they had seven goats. All the goats had one bow opn there each horn. Each goat was carrying one bag of corn. How many things are there in all?*
Solve the problem. Show your work in the space below. *7* *14* *2* *7* *30*
Describe how you solved the problem in your own words in the space below. You may use pictures in addition to words to describe your problem-solving process. *I add 7 goats + 14 ribbons. + two people. + 7 sacks of corn.*

Example 3

Student M0070 earned a score of 2 on each dimension for a total score of 8.

Write the word problem in your own words in the space below.

you have to ad sacks and keep adding everything up

Solve the problem. Show your work in the space below.

sacks 7
ribbons 7
people 2
goats 14
30

Describe how you solved the problem in your own words in the space below. You may use pictures in addition to words to describe your problem-solving process.

(picture of seven sacks)

(picture of one goat with the expression + 6)

(picture of two people)

(picture of bow with the expression 6+)

Example 4

Student M0118 earned 1 point on each of the dimensions for a total score of 4.

Write the word problem in your own words in the space below. *Ther wer 7 goats me he sack of corn and there own the rode*
Solve the problem. Show your work in the space below. *I am cfusde*
Describe how you solved the problem in your own words in the space below. You may use pictures in addition to words to describe your problem-solving process. *I am cfusd*

11

Toys in Space

This chapter includes the *Toys in Space* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form and are available for reprint and use.

Development History

This assessment was also developed by the project staff rather than a school team. Its original is somewhat less-well-grounded in any experience base in assessment or instruction. The idea for *Toys in Space* was sparked by an announcement on the radio about a live interactive communication between the astronauts demonstrating principles of physics using toys and their home-town elementary schools. The idea of students interacting with astronauts via tele- and videocommunication technologies focusing on science learning was too rich an opportunity to miss. The project team felt that there was a high likelihood that some meaningful assessment task for the portfolio could be generated from this interactive session. Furthermore, the diversity of stimuli used in the aforementioned tasks did not include any that were technologically sophisticated. Thus, from the perspective of diversity in stimulus, the notion of using the live transmission between earth and space was enticing.

After numerous telephone calls to NASA, both to the teacher center at Cape Kennedy and Houston, as well as to Dr. Carolyn Sumners, director of the Toys in Space program at NASA, the project staff learned about the instructional underpinnings of the transmission and the gist of the interaction. The next step was to capture the transmission. This was done with the permission of NASA (readily available to the public). We taped the live transmission rather than wait for NASA to include this transmission in their *Lift-off to Learning* series because we wanted to move quickly to see if the transmission afforded the project with meaningful science content to use within the context of assessment.

Once the video was obtained, we asked an expert science education consultant to review the videotape. The purpose of this review was to determine whether or not there was meaningful science content presented during the transmission. If this was the case, the second request was that some recommendation be made about which of the four astronaut presentations would be most appropriate to the student population addressed in the Authentic Assessment for Multiple Users (AAMU) student (grades 3 through 6). The results of this expert review were quite positive. The specific meaningful and important science content captured in this entire video are reported in the description included in the Teacher Instructions. The recommendation of where to focus was on the astronaut playing with the wind-up car and track.

Toys in Space was designed as an assessment entry for the portfolio that builds upon Science Observation and the instructional work of Ellen Doris. The student was presented the stimulus video. There is a series of seven questions to which the astronaut responds. Prior to his response to the seventh question, the videotape is stopped and the student is asked to predict the answer to that question. After a brief pause enabling the student to write and draw this prediction, the video is resumed. The student is then asked to observe and record (through writing and drawing) what actually happened. Sixth-grade students are then asked to analyze the difference (if any) between what he/she predicted and what actually happened. Finally, all students are asked to write a question they would like to pose to the astronaut if given the opportunity -- an extension question.

This assessment is unique in that instructional support was provided to the teachers in advance of the assessment through materials from NASA and the Lift-off to Learning series. The physics content was literally imported into these classrooms with the teachers' permission because not only was physics not part of the three through sixth curriculum but the participating teachers were not at all prepared to teach this content. They were, however, sufficiently intrigued with the notion of using videotapes as the content stimulus to test this technique. Furthermore, they were compelled to find diverse documentation strategies to describe what student think, know, and can do in both science and mathematics.

The project goals addressed across these four to five products (depending upon the grade level of the students) were "reflective thinkers and self-evaluators," "creative and strategic thinkers," "self-directed learners," "effective communicator," and "responsible global citizens."

Relative to the dimensions of time, content complexity, stimulus complexity, and response complexity, it scores as follows:

Attribute	Score
Instructional Time	Five to eight days of instruction
Content Complexity	Complex
Stimulus Complexity	Complex
Response Complexity	Complex

Compared with regard to the dimensions of time, content complexity, stimulus complexity, and response complexity present in any of the previously described assessments, *Toys in Space* is the most complex and sophisticated both in terms of content and presentation. The mode of communication is writing and drawing.

TOYS IN SPACE

ASSESSMENT

Toys In Space: *Teacher Instructions*

TOYS IN SPACE ASSESSMENT

Overview

The videotape entitled *Space Shuttle Physics Experiments (With Toys)* presents a forty-minute science lesson conducted from space by the astronauts on the space shuttle Endeavor during the January 1993 mission. The crew of the Endeavor included John Casper (mission commander), and Mario Runco, Greg Harbaugh, Susan Helms, and Don McMonagle (mission specialists). The focus of the lesson was on how some typical children's toys behave in space. During the lesson, each of the astronauts takes questions about one specific toy from students currently enrolled in the elementary school the astronaut attended. After each question, the astronaut investigates the question, thereby providing the students with the opportunity to observe the answer. The astronaut then explains what was observed.

Concepts

The physical science concepts demonstrated during the entire lesson include:

- free fall or microgravity
- friction
- momentum
- conservation of momentum
- centripetal force
- work
- energy
- contact forces and action-at-a-distance forces
- Newton's Laws of Motion
  - First Law: Law of Inertia
  - Second Law: F = ma
  - Third Law: For every action there is an equal and opposite action.
- magnetism
- angular momentum
- gravity
- frames of reference
- Law of Conservation of Momentum

Toys In Space: *Teacher Instructions*

Purpose

The purpose of this assessment is to document students' abilities to "do what scientists do" (following E. Doris, *Doing What Scientists Do: Children Learn to Investigate Their World, Heinemann*, 1991).

At third-, fourth-, and fifth-grade levels, the students are asked to listen and watch, to predict, and then to observe and record, both pictorially and verbally, what they see.

At sixth-grade, the students are asked to listen and watch, to predict, to observe and record, both pictorially and verbally, what they see, and then to analyze their prediction in relation to what actually happened during the investigation.

In addition, at each grade level the students will be asked whether they could replicate the behavior of the toy on Earth, and they are to write one question about the toy's behavior in space that they would like to have the astronaut answer.

Setup

The track consists of two 180^0 arcs. These tracks should be assembled prior to the pre-assessment activity. Each team of students should have one car and one 360^0 track. For the assessment, each student should have the assessment worksheet.

In addition, the classroom teacher will need a video player with a counter on it. The video monitor must be in clear view of all students.

Time

The pre-assessment activity should be spread across several class periods prior to the day on which the assessment is given. NASA has several instructional videos which can provide the instruction/guidance required for understanding microgravity (free fall) and friction. Specifically, *Space Basics* and *Newton in Space* used in that order provide ample content to precede the assessment. In addition, in order to spark an interest in the shuttle Endeavor and space in general, the entire pre-assessment activity would be well served by beginning with the video *Endeavor Then & Now*. Furthermore, there are some materials for the teacher's use in the *Toys in Space* and the *Liftoff to Learning* publications from NASA. The pre-assessment activities (including the orientation to the toy car and track, as discussed below, plus the orientation to microgravity and space) may take between one and two class periods. It is important that the actual assessment activity take place within a single, uninterrupted block of time within a class period.

Pre-Assessment Activity

The purpose of the assessment is to document student reasoning about how a toy behaves in a controlled environment in space, far removed from any real-world experience that the students may have had. It is important that each student have an opportunity to investigate how the toy behaves in the classroom environment. For this reason, the students should be given the toy and track for up to one class period the day before the assessment is scheduled. The students should be encouraged to explore and to see how the toy works (how it works in general and, specifically, how it works on the track). This activity should be guided only by the children's curiosity and by the restriction that no harm should come either to the toy and track or to individuals!

The students should work together in teams of two or three during this pre-assessment investigation.

Materials

Each student needs a WORKSHEET and a pencil. The teacher/assessor needs the videotape, a videotape player with a counter, and a video monitor. All students should have a direct view of the video monitor.

Lesson Concepts

The concepts covered during the car lesson include:

- Inertia.
- Contact forces.
- Law of Conservation of Momentum *(acceleration = velocity squared/radius).*
- Free fall or microgravity.
- Gravity.

Instructions

Listen carefully while I explain what you are going to do today. Think about the toy car and the fun you had playing with it. That same car and track went into space with the astronauts on the space shuttle Endeavor in January, 1993. The astronauts played with the car and track just as you did.

Today, we are going to watch a videotape about the space shuttle Endeavor and what the atmosphere is like in the shuttle in space. The videotape also shows Astronaut Mario Runco investigating how the toy car behaves in space.

Third-, fourth-, and fifth-grade students:

Astronaut Runco will listen to a question about the toy car from a student in the same elementary school where he went to school. Then he will attempt to answer the question by demonstrating how the toy car behaves. Your task is to listen and watch carefully. Think about the question that the student asks. You will be asked to write down your prediction. After you have made your prediction, we will then continue the videotape and see what actually happened. Next, you will be asked to record what you observed. You will draw a picture of what the car did, and you will write about what you saw.

Let's begin. (Start the videotape at 3:46. Stop the videotape at 18:08 immediately after the question, "*Will the car jump its hole if the track is opened?*")

The question is, *"Will the car jump its hole if the track is opened?"*

What do you think the car will do? Write down your prediction, and include the reason for your prediction. (Pause)

Now watch and see what happens. (Stop the tape at the end of Astronaut Runco's demonstration.) You should now draw a picture of what you saw. When you have finished that picture, write about what you noticed.

Next, explain why you think the car did what it did.

Finally, think about a question that you would like answered about how the toy car would behave in space. Write that question on the last page in the Student Worksheet.

Toys In Space: *Teacher Instructions*
 Project funded by the National Science Foundation.

Sixth-grade students:

Astronaut Runco will listen to a question about the toy car from a student in the same elementary school where he went to school. Then he will attempt to answer the question by demonstrating how the toy car behaves. Your task is to listen and watch carefully. Think about the question that the student asks. You will be asked to write down your prediction. After you have made your prediction,we will then continue the videotape and see what actually happened. Next, you will be asked to record what you observed. You will draw a picture of what you saw and you will write about what you saw.

Let's begin. (Start the videotape at 3:46. Stop the videotape at 18:08 immediately after the question, *"Will the car jump its hole if the track is opened?"*

The question is, *"Will the car jump its hole if the track is opened?"*

What do you think will happen? Write down your prediction, and include the reason for your prediction. (Pause)

Now watch and see what happens. (Stop the tape at the end of Astronaut Runco's demonstration.) You should now draw a picture of what you saw. When you have finished that picture, write about what you saw.

Then, think about how your prediction did or did not describe what really happened. In the place marked "Summary" write an explanation of why you did or did not accurately predict what happened.

Next, explain why you think the car did what it did.

Then, explain why your prediction and what actually happened were either the *same* or *different.*

Finally, think about a question that you would like answered about how the toy car would behave in space. Write that question on the last page of the Student Worksheet.

Toys in Space Assessment
Student Response Worksheet

Student Name: _____

School Name: _____

System Name: _____

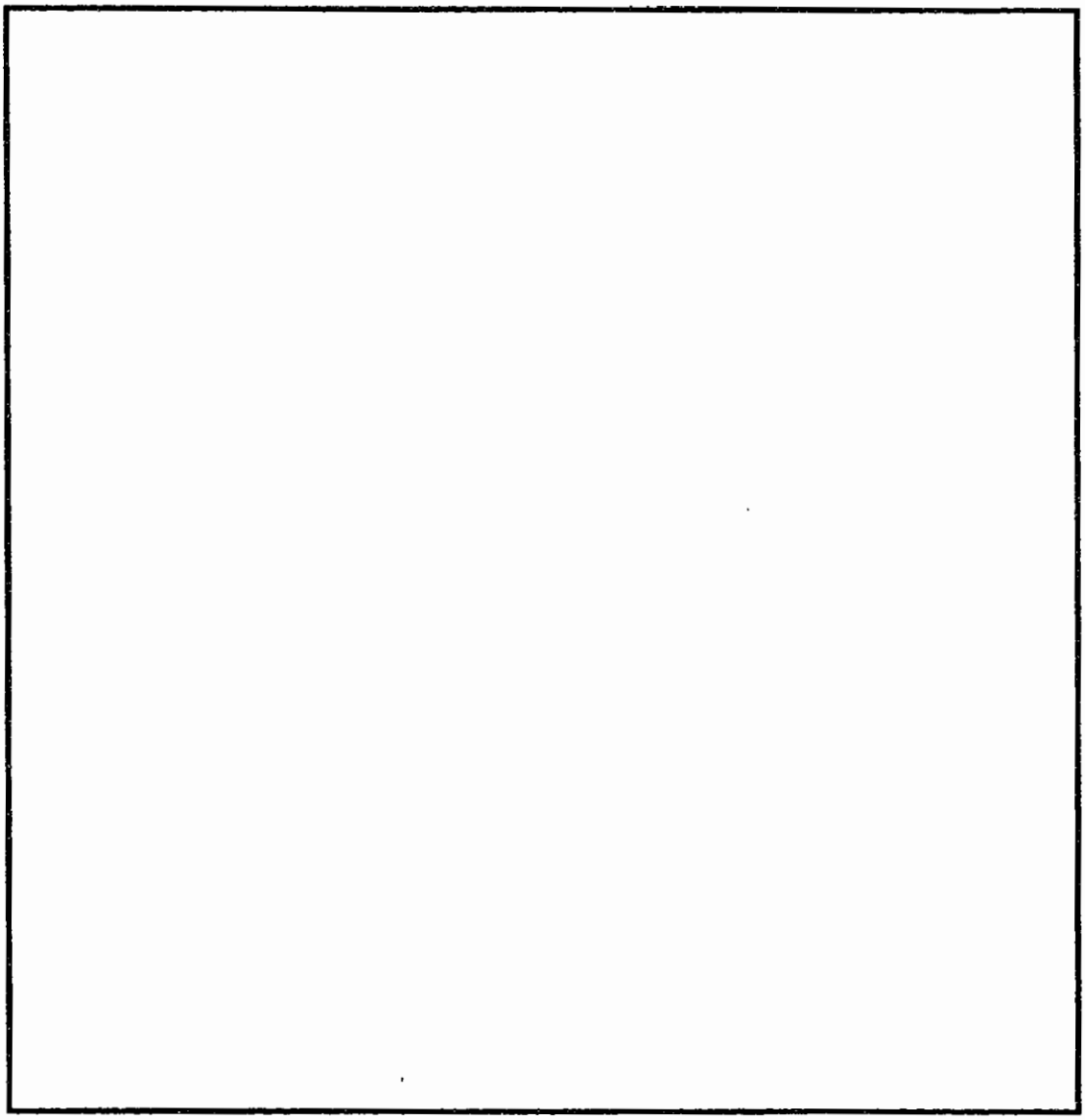
Teacher: _____

Grade: _____ Date _____

What do you think the car will do?[1] Why?

[1] Ellen Doris, *Doing what scientists do: children learn to investigate their world, Heinemann, 1991.* Permission granted from Heinemann to reproduce this format.

Draw a picture of what the car did.



What did you notice?

Summary (sixth-grade only)

Why do you think the car did what it did?

In your classroom, could you get the car to do the same thing as the astronaut was able to get it to do? Explain. (sixth-grade only)

Toys In Space: *Student Response Worksheet*
 Project funded by the National Science Foundation.

What question about the toy car would you like to ask the astronaut?

The scoring guide for *Toys in Space* is analytic. Thus, here is the second assessment entry for the portfolio that builds upon the work of Ellen Doris and her book "Doing What Scientists Do" that complements *Science Observation* (also built upon this same foundation) but that offers an analytic scoring approach. It is important to notice that with Toys in Space, the information that can be reported to stakeholders is not directly comparable to that available through Science Observation. In Toys in Space, the information available is for the discrete elements of interest (prediction, drawing, narration, analysis, and questioning skills) Within each focus area, the possible points range from 0 to 4. A composite score could be built by adding the discrete scores.

Scoring Guide: Toys in Space

	Prediction (What do you think the car will do? Why?)
0	No attempt, or a prediction unrelated to the experiment
1	Prediction with no rationale
2	Prediction with a rationale
3	Accurate prediction with a rationale
4	Exemplary (includes accurate prediction and appropriate rationale for the variables involved)
	Drawing (Draw a picture of what the car did.)
0	No attempt
1	Attempt (includes car and track)
2	Appropriate (includes car and track, shows motion)
3	Accurate representation (includes car and track, shows motion, and demonstrates free flight of car away from the track)
	Narrative (What did you notice? Why do you think the car did what it did?)
0	No attempt
1	Attempt (mentions car and track)
2	Appropriate (includes car and track, mentions motion)
3	Accurate representation (includes car and track, mentions motion, and discusses free flight of car away from the track)
4	Exemplary ((includes car and track, shows motion, discusses free flight of car away from the track, includes reference to term or concept of gravity, microgravity, etc.)
	Contrast of space with Earth (sixth-grade students only) (In your classroom...? Explain.)
0	No attempt
1	Attempt
2	Appropriate (but no evidence of extension of scientific concepts or principles)
3	Accurate Attempt (some evidence of extension of scientific concepts or principles)
4	Exemplary (demonstrates evidence of extension of scientific concepts or principles; indicates logical "next step" in the investigation)
	Question (What question about the toy car would you like to ask the astronaut?)
0	No attempt
1	Attempt (but not appropriate, given task)
2	Appropriate (but no evidence of extension of scientific concepts or principles)
3	Accurate attempt (some evidence of extension of scientific concepts or principles)
4	Exemplary (demonstrates evidence of extension of scientific concepts or principles, indicates logical "next step" in the investigation)
	NOTE: Summary question is not scored.

Toys In Space: *Scoring Guide*
 Project funded by the National Science Foundation.

Examples of Scored Student Work

As presented on page 178, the scoring guide for *Toys in Space* is also analytic with five scoring dimensions. Relative to the dimensions of the analytic scoring guide, they relate to the questions on the assessment worksheet as follows:

Drawing	**Draw a picture of what the car did**
Narrative	**What did you notice?**
	Why do you think the car did what it did?
Prediction	**What do you think the car will do?**
Question	**What question...ask the astronaut?**
Contrast	**In your classroom....Explain**

As with typical analytic scoring guides, there are multiple ways to generate certain composite scores. The examples included in this chapter are intended to demonstrate that character as well as to demonstrate the range of thinking evident in third, fourth, and fifth grade students in this very complex content area of physics. The task itself has potential for use outside of this particular content area -- not only in science but in other disciplines as well.

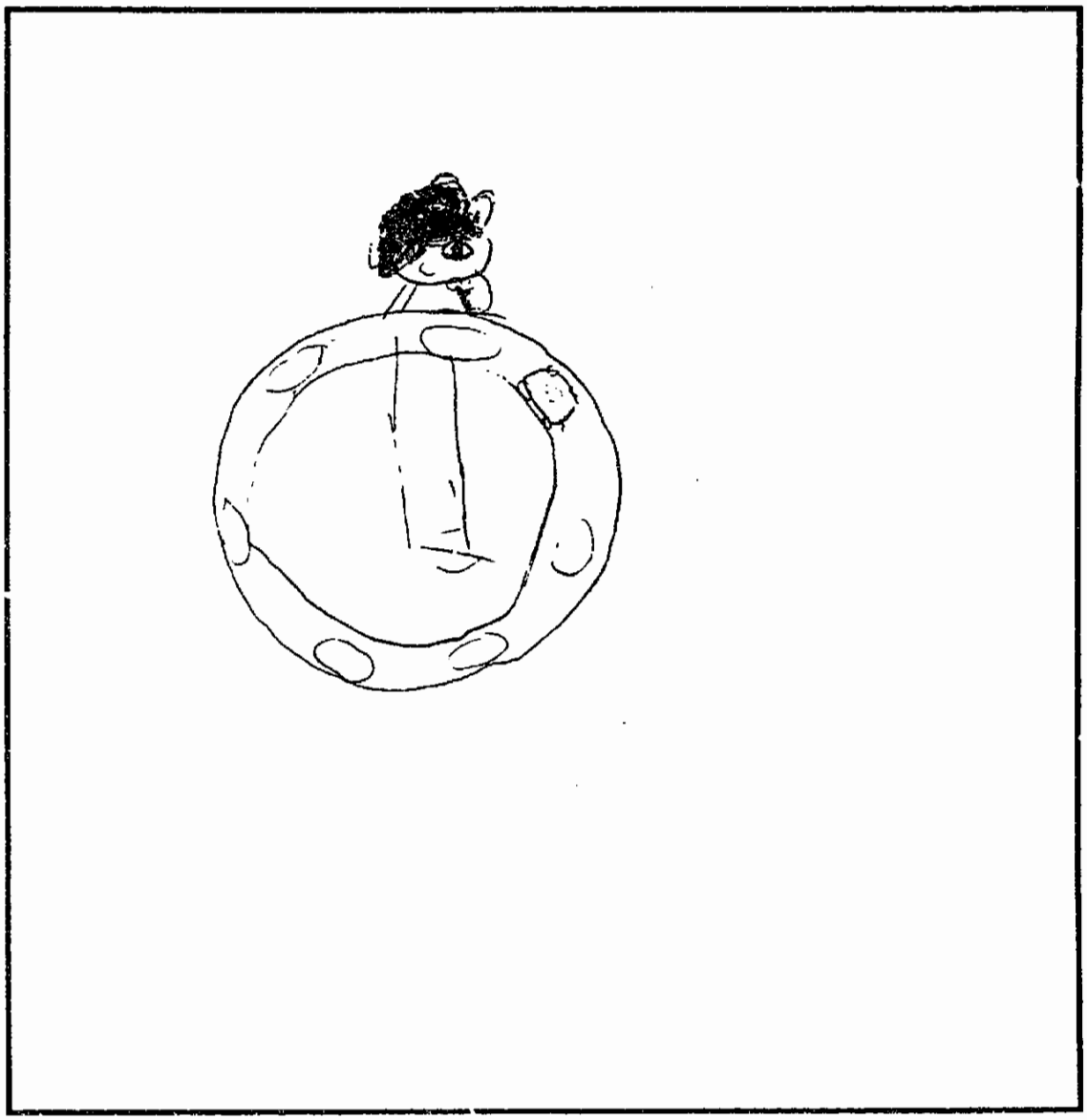
Example 1

Student C0055 received a composite score of 14 ("4" for Prediction, "2" for Drawing, "4" for Narrative, "0" for Contrast, and "4" for Question).

What do you think the car will do?

I think it going to fly out fast and because of no gravity because of microforce.

Draw a picture of what the car did. Student C0055



What did you notice?

The car spined from behind the wheels were going fast but there was nothing to prevent the wheels. The contact with his hands are stronger. The top of the car couldn't move. There was a lot of mass on the top of the car after the car goes fast, it slows down and goes of the track its beause of microgravity. The Force that made the car its wheels down by gravity centripetal force made the car go around.

Why do you think the car did what it did?

The car went of the track because he opened the track and there was no gravity in the air. the force of gravity was pulling the ship which made it orbit around the earth. it was stronger than the gravity moving the car.

Summary

(no response)

In your classroom, could you get the car to do the same thing as the astronaut was able to get it to do? Explain.

(no response)

What question about the toy car would you like to ask the astronaut?

1. *Why is there no gravity on space?*
2. *Can the cars go around when the men isn't holding the track?*
3. *How is it that there is a such thing as micro gravity?*

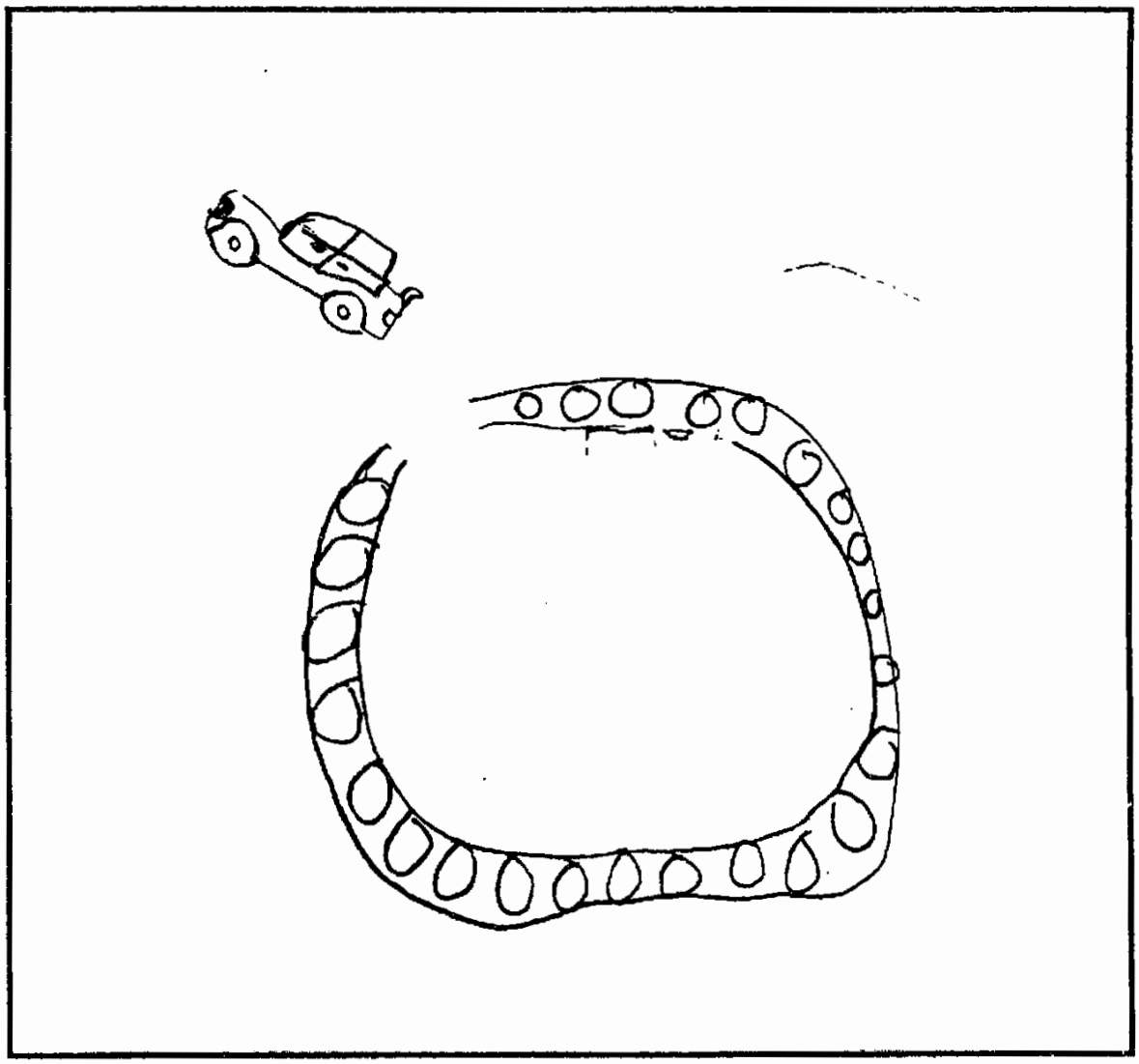
Example 2

Student R0003 received a composite score of 13 ("1" for Prediction, "3" for Drawing, "4" for Narrative, "3" for Contrast, and "2" for Question).

What do you think the car will do?

It will shoot out straight in the direction of.

Draw a picture of what the car did.



What did you notice?

On earth it would fall. In space it shoots straight up.

Why do you think the car did what it did?

In space the car has no gravity.

Summary

(no response)

In your classroom, could you get the car to do the same thing as the astronaut was able to get it to do? Explain.

No because with gravity on earth the car won't float.

What question about the toy car would you like to ask the astronaut?

If you had an airplane would it fly straight?

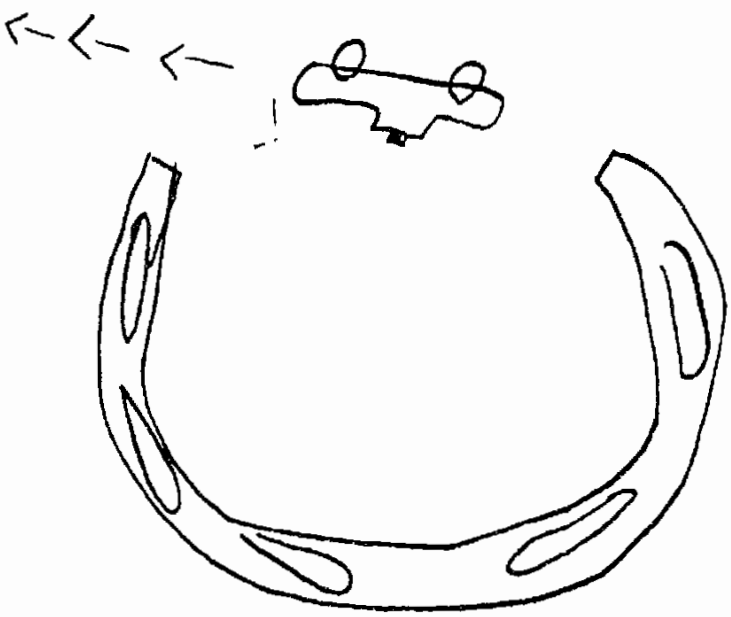
Example 3

Student M0117 received a composite score of 11 ("1" for Prediction, "3" for Drawing, "3" for Narrative, "2" for Contrast, and "2" for Question).

What do you think the car will do?

It willl go through the hole.

Draw a picture of what the car did.



What did you notice?

I notice when he opened the track the car went straight off the track.

Why do you think the car did what it did?

I think the car went off the track because there was nothing the wheels could go on thats why it went off the track.

Summary

(no response)

In your classroom, could you get the car to do the same thing as the astronaut was able to get it to do? Explain.

Yes our car went out of the tracks when it got to the hole in it. It did not jump the track. The direction it went in when it got out of the track was diaganle and then down.

What question about the toy car would you like to ask the astronaut?

What would the car do on the track if you did not whinned it up?

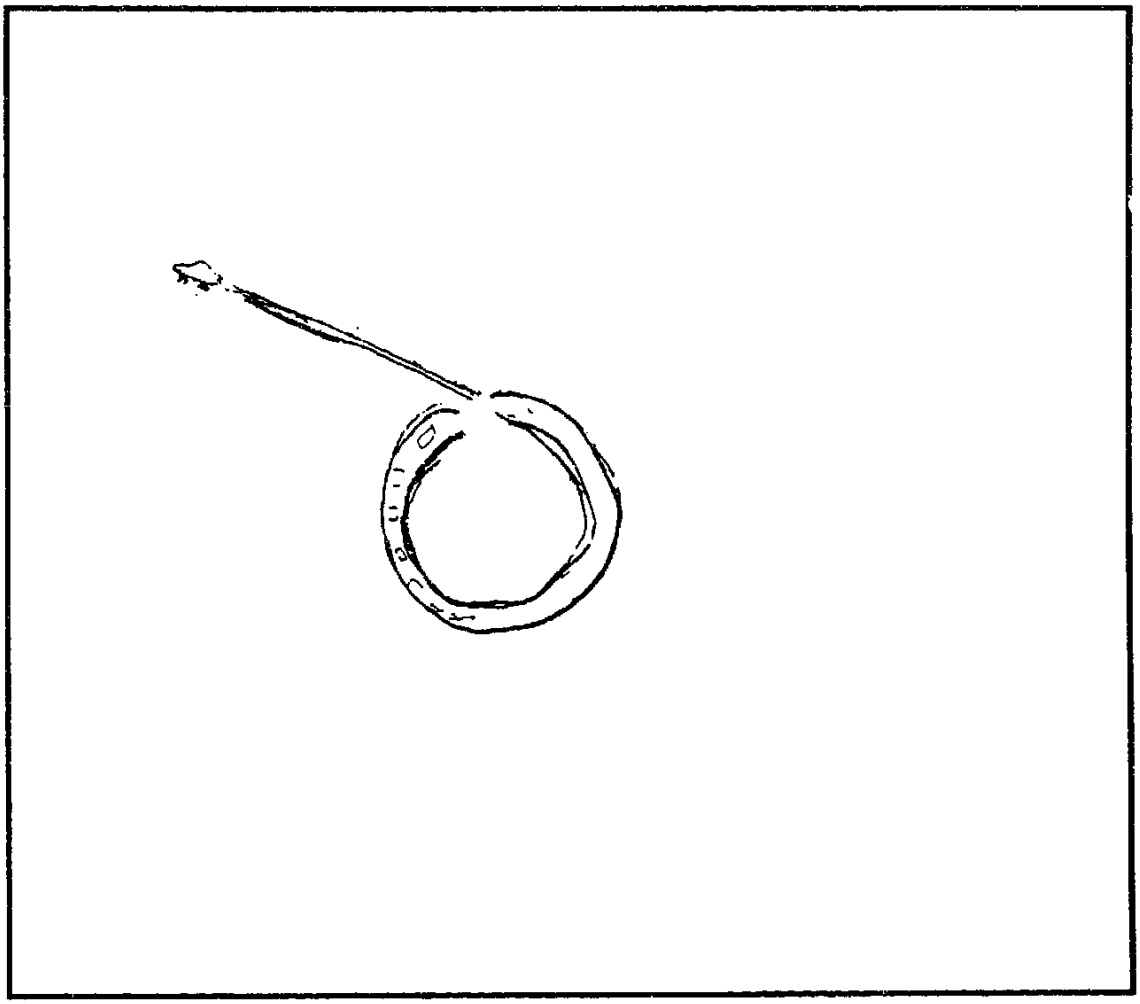
Example 4

Student C0080 received a composite score of 9 ("2" for Prediction, "3" for Drawing, "2" for Narrative, "0" for Contrast, and "2" for Question).

What do you think the car will do?

My prediction is that the car goes up because when he didn't have it on the track the front went up but this time the track while open and it while go the angle the track is facing.

Draw a picture of what the car did.



What did you notice?

When he let the car go its front end went up because the wheels were spining real fast

Why do you think the car did what it did?

The reason the car would go up is because the front is angled up because the back is heavy and the reason it would keep going is because the whells are spining so fast. When the whilles stop it stops moveing. Not all the way though.

Summary

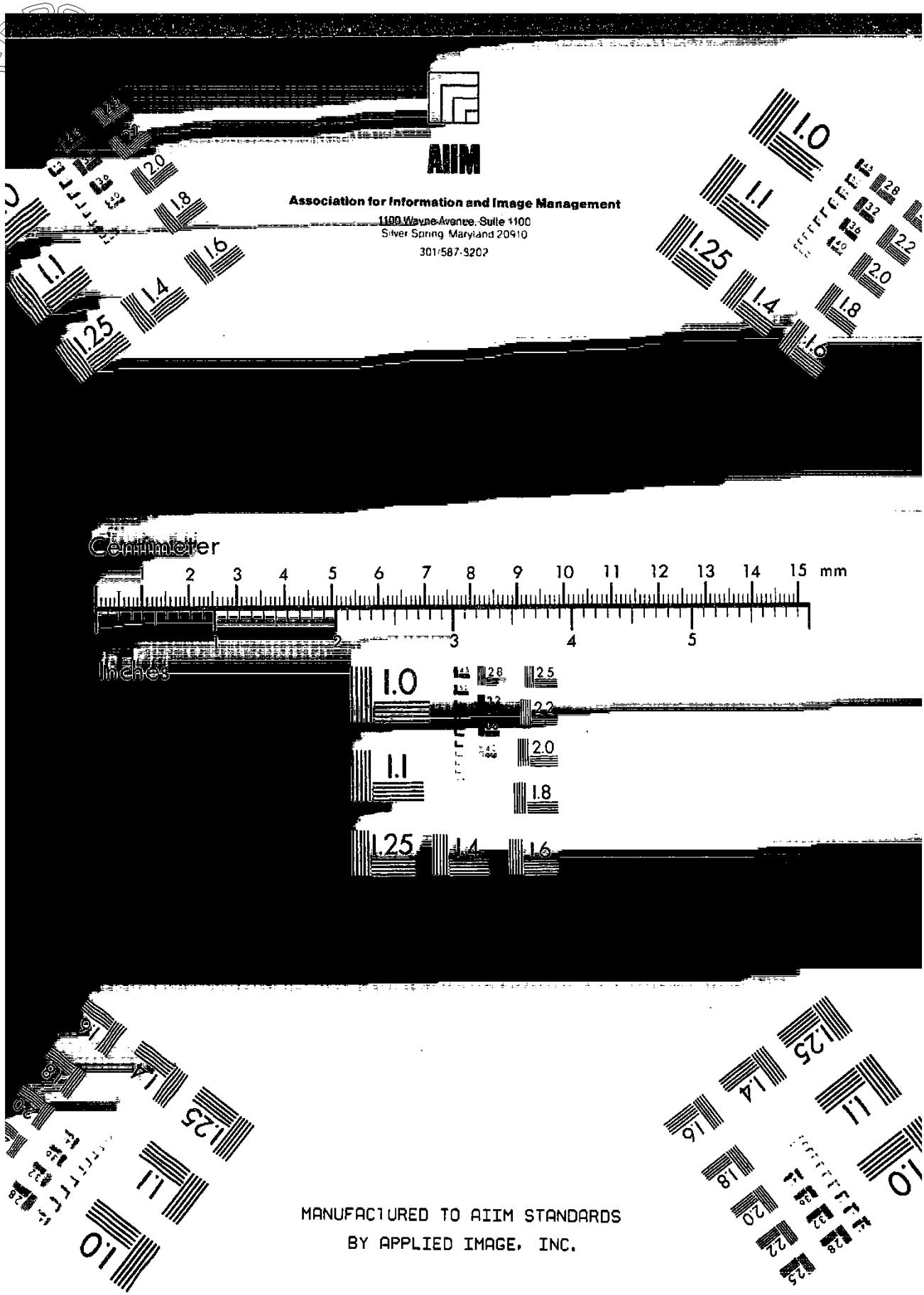
(no response)

In your classroom, could you get the car to do the same thing as the astronaut was able to get it to do? Explain.

(no response)

What question about the toy car would you like to ask the astronaut?

1. *Why doens't your hair float and the car does*
2. *How do people float when they are in space because they have 90% gravity?*



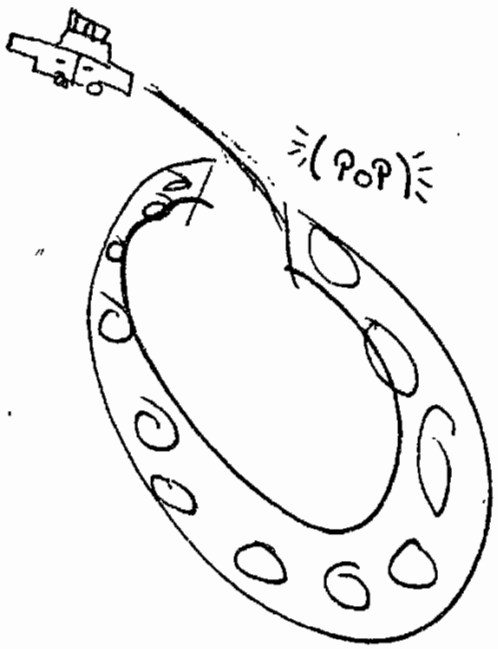
Example 5

Student R0065 received a composite score of 6 ("3" for Prediction, "3" for Drawing, "0" for Narrative, "0" for Contrast, and "0" for Question).

What do you think the car will do?

No because there's to much gravity and the car will start to float away.

Draw a picture of what the car did.



What did you notice?

Well i noticed that

Why do you think the car did what it did?

(no response)

Summary

(no response)

In your classroom, could you get the car to do the same thing as the astronaut was able to get it to do? Explain.

(no response)

What question about the toy car would you like to ask the astronaut?

(no response)

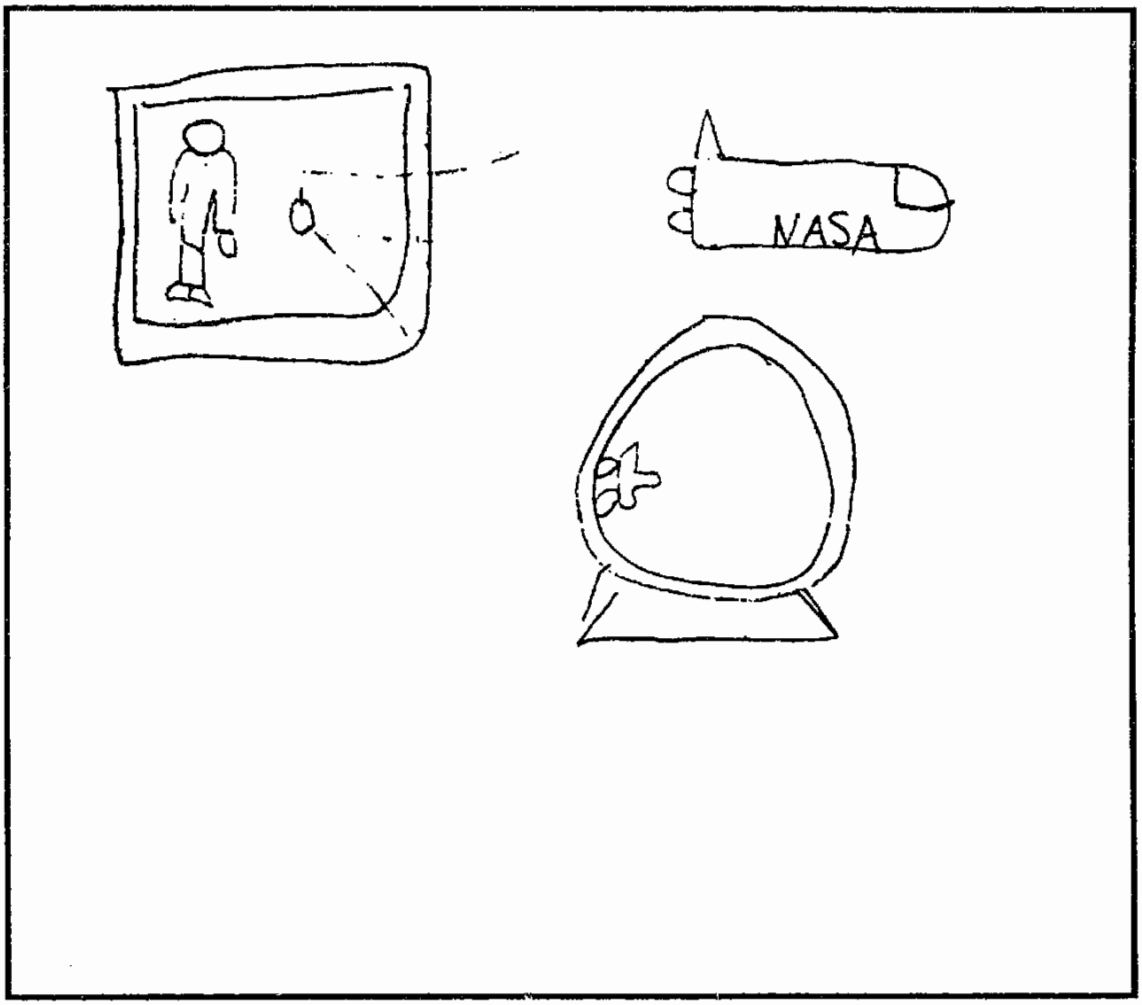
Example 6

Student C0062 also received a composite score of 6 but quite differently than the performance of Student R0065 ("1" for Prediction, "1" for Drawing, "2" for Narrative, "0" for Contrast, and "2" for Question).

What do you think the car will do?

My prediction is that the car is going to float.

Draw a picture of what the car did.



What did you notice?

I noticed that the race car go alot faster in space than in earth.

Why do you think the car did what it did?

My analysis is that the car is going to gloat because of gravity when it got up to the top the car could not keep going because ther was a hole.

Summary

(no response)

In your classroom, could you get the car to do the same thing as the astronaut was able to get it to do? Explain.

(no response)

What question about the toy car would you like to ask the astronaut?

If you shake the track with the car on it what would happen.

12

Interview Assessment

This chapter includes the *Interview Assessment* task, teacher directions, scoring guides and support materials as well as chronicles of how this task came to be and how it should be presented. This task and its ancillary materials are presented in camera-ready form and are available for reprint and use.

Development History

From the moment the AAUM project began and the teachers/researchers met for the first time, there was a compelling interest in interviewing the students about what they knew, had learned, and needed to learn. Countless teachers reported that they were certain that if they could just talk with each student in depth about what they had or had not learner, that the information gleaned would be invaluable. The Marietta school team pursued this dream. They developed an *Interview Assessment* that initially called for the videotaping of all interviews for the purpose of scoring each videotape in a systematic and reliable manner by two or more raters.

The *Interview Assessment* consisted initially of the students being questioned in detail at the conclusion of a unit or set of lessons. This conversation was then videotaped for future scoring. The real limitations of this plan became clear early in a small-scale tryout. Specifically, the burden of time required to view and score each twenty to thirty minute videotaped conversation became prohibitive. This time burden far exceeded the value of the information provided. Thus, the teachers/researchers rethought their assessment with specific changes to model itself. The teachers conceded that somehow the score or credit for different questions needed to be recorded at the time of the interview rather than later during a re-viewing. So, they essentially redesigned this assessment from the perspective of a behavioral checklist. The content of this revised *Interview Assessment* is mathematics but some applications/problems based in science content would also be appropriate.

The student is presented a problem and asked to both solve it and to reflect upon the strategies employed for solving it. The information desired from the students focus on whether or not a student can restate the problem presented, whether or not they can explain the procedures employed and whether or not the results are reported reasonable and correctly. The project goals addressed across these four focus areas "creative and strategic thinkers," "self-directed learners," "effective communicator," and "strategic and reflective thinkers." Relative to the dimensions of time, content complexity, stimulus complexity, and response complexity, it scores as follows:

Attribute	Score
Instructional Time	Lesson specific (probably 3 to 5 days)
Content Complexity	Variable
Stimulus Complexity	Variable
Response Complexity	Variable

Compared with the dimensions of time, content complexity, stimulus complexity, and response complexity present in any of the previously described assessments, *Interview Assessment* is more narrow and limited in terms of learning sampled. The evidence of student learning is closely tied in terms of the complexity to the problem selected by the teacher to serve as a stimulus. The mode of communication is speaking and some writing.

Teacher Instructions

INTERVIEW ASSESSMENT

Teacher Instructions: Interview—Mathematics Problem Solving

Overview

Students are presented with a math word problem that they are asked to solve, showing their work on a sheet of paper. Students are then interviewed to assess their mathematical skills and process strategies for solving the problem.

Purpose

The task was designed to evaluate students' ability to solve math word problems and orally explain their strategies. Students are evaluated on their:

1) Ability to explain the problem and the essential elements.

2) Ability to explain the solution strategy (the steps used in solving the problem).

3) Ability to reach an accurate solution, given the problem, and supported by the data from the work they present.

Planning

√ **When you plan the math context/problem to be used for this evaluation the following must be considered:**

1) The problem should be a multiple-step word problem of sufficient complexity so that problem-solving skills can be evaluated.

2) If the same problem is to be used for all of the students in a classroom, it may be necessary to interview them outside of the classroom to avoid the transfer of information about how the problem might be solved from the interviewee's being overheard. If different problems are to be used, they must be of the same complexity and require the same mathematical skills if comparisons are to be made across students.

3) Each student should be recorded on audiotape. Make sure you identify the person being taped for reference during scoring.

4) Prior to administering the task, have the students practice "talking about" how they solve problems. Model an interview.

5) It may be helpful to note the level of probing necessary for each student during the interview. Have students explain all of the work that appears on their worksheets.

6) Conduct the interviews as soon after the problem is introduced as is reasonable. Students may forget some of the processes and reasoning used while solving the problem if a great deal of time elapses.

7) After presenting the word problem to students, explain that they will be asked to restate the problem in their own words and will later be interviewed to explain the steps they took to find the solution. Remind them to "show" their work and the steps they took to solve the problem on their worksheet (which will be used as a reference during the interview). Encourage them to make notations that might clarify or support their work.

√ **Attach a copy of the word problem and the student's worksheet to the Interview Scoring Record.**

Note. *The Interview Scoring Record may be used to score the student at the time of the interview, unless independent readers are required. Also, there is a place to record comments during the interview on the Interview Scoring Record.*

Interview
Student Worksheet

Student Name: _____

School Name: _____

System Name: _____

Teacher Name: _____

Grade Level: _____ Date: _____

Interview: *Student Response Worksheet*
 Project funded by the National Science Foundation.

Interview Scoring Record

Student Name: _____

Question # 1. ***Tell me, in your own words, what the problem was. What were you supposed to do?***

Restates the task		
	RUBRIC	COMMENTS
0	Unable to restate the problem	
1	Restates the problem with some errors or incomplete information	
2	Restates the problem accurately	

Question # 2. ***Tell me how you solved the problem. What steps did you use from beginning to end?***

Note. You want to find out the reasons for the steps in the procedure and the information or elements necessary to solve the problem. Therefore, you may need the following probes: Tell me why that's important. Tell me what you did to find that out. The information provided on the sheet that students are using to "show their work" will be helpful in guiding the probing.

Explains Procedures		
	RUBRIC	COMMENTS
0	Procedures are not workable.	
1	Some of the procedures described are workable *and* some of the necessary elements are included.	
2	Procedures described are workable *and* most necessary elements are included. The procedures, if carried out correctly, would lead to a solution.	
Reports Results		
0	Solution is incorrect; it is evident from the data (or lack of data) that the student did not know how to solve the problem.	
1	Solution is incorrect because of minor errors in calculations, *but* there is evidence of support with data for the solution **or** the solution is correct, but there is no evidence of support with data for the solution.	
2	Solution is correct *and* the student supports the solution with data.	

Interview: *Scoring Guide*
 Project funded by the National Science Foundation.

The scoring guide for *Interview Assessment* is analytic in like vein with *Toys in Space*. The focus elements are: (1) restates the task, (2) explains procedures, and (3) reports results. Within each of these three areas of focus the score points possible range from 0 to 2. No specific "off task" identification is offered. As with *Toys in Space* or *Retelling*, the discrete analytic scores may easily be aggregated to present a total or overall score. However, once that is done, the information yield from this assessment tool is seriously limited.

A Holistic View of The Assessment Entries for the Portfolio

As the participants in the AAMU project began to develop assessment entries for the portfolio, one goal was to include tasks/activities that were diverse in their structure, format, and student learning goals. In addition, we strove to include tasks/activities that were different with respect to the attributes of complexity described on pages 33-34. As reported in Table 12.1, considerable variation in each of these attributes was evident from the development perspective. Whether or not the evidence elicited from students also demonstrates this variation is a question that should be examined in each application.

Table 12.1 Attributes of Assessment Tasks

TASK	ATTRIBUTE			
	Instructional Time	Content Complexity	Stimulus Complexity	Response Complexity
Letter Writing	Typical Unit (3-6 weeks)	Variable	Simple	Variable
Science Observation	Variable	Variable	Variable	Variable
Problem Solving	Unspecified since the content covered may span an entire year of science study	Variable	Variable	Relatively limited by the implicit paradigm suggested by the Student Worksheet and its resemblance to a standard lab report form
Comparison of Experiments	Four to eight weeks (multiple units/lessons)	Complex	Complex	Complex
Continuum of Progress	Multiple units over large periods of time (quarter, semester, year)	Complex and Variable	Complex and Variable	Ranges from simple to complex over the questions in the assessment
Retelling	Unspecified	Fixed	Fixed	Variable depending upon the student's mathematics sophistication
Toys in Space	Five to eight days of instruction	Complex	Complex	Complex
Interview	Lesson specific (probably three to five days)	Variable	Variable	Variable

At issue here is whether the portfolio has meaning as more than simply the collection of individual points of evidence. If portfolios are simply databases in which points of evidence are stored, then the meaning of portfolio is limited in its ability to better describe student learning than conventional collections of evidence. On the other hand, if by the purposeful collection of many and varied points of evidence, the picture of each student as a learner is enhanced and amplified, the assessment portfolio makes a significant contribution to teaching and learning.

Given the paradigm of the assessment portfolio that emerged from this project, the collection of evidence about student learning has, at its heart, a common structured core of meaningful work. This core must in its totality reflect the intellectual work of the discipline, both content and process. This core is the basis for trend analysis and other forms of aggregate study. Then, around this core are the less than systematic points of evidence, those report mechanisms that inform about work preferences, areas of special interest, areas of particularly impressive or challenging growth, and so forth. Also included in this non-systematic or more idiosyncratic entries in the portfolio might be teacher and parent descriptors of change in the student relative to the discipline(s).

One of the major criticisms of traditional multiple-choice tests voiced by teachers and administrators is that the test-taking work itself is not viewed as meaningful. In portfolio assessment, we run the risk of suffering a like disposition. Just by moving from traditional testing to portfolio assessment or to performance-based assessment, we may still present an image of irrelevant or useless work. We must not use portfolios as a way to be less informative about what expected work is in a discipline. We must not use phrases like reflection to escape the important accountability for each student that answers the question; what does each student think and know, and how do they exhibit the ability to do science and mathematics and every other discipline.

The *Interview Assessment* 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 difficulty in obtaining high quality video tape, the developers decided to redesign the interview to not rely on live documentation. That was discussed in Chapter 6. No exemplars were identified and no large scale scoring occurred of this assessment. The *Interview Assessment* was tested however, and four raters were able to score the interviews consistently. It is included here as a draft second edition task.

13

Scoring Guides and Implementation

This chapter describes the process of developing scoring guides, the process of scoring the student responses and the agreement data among raters. Scoring guides were developed after the initial tasks were tested through small scale try-outs. These try-outs enabled the task developers to refine the scoring guides in light of "real" student responses. However, the question of what constitutes quality work for all students remains an important question that justifies continual re-examination. This chapter includes a discussion of how training materials for scoring were developed and how training for scoring was conducted. Finally, the bottom line data is interpreted relative to assessment task quality and potential need for modification.

Scoring Guide Development

Scoring guides must contain information that amplifies understanding of the descriptive performance levels. This information is specific pieces of student work on the assessment being scored. These pieces of student work represent "exemplars" (best examples) of each score point in the scoring guide. Often these exemplars come from small-scale try-outs. Sometimes they are initially generated by the assessment developer based on research. Often they are a combination of student-generated responses to the assessment task and modifications of student responses to particularly illustrate a score point. The clarity of the relationship between the exemplar and the score point is critical if the exemplars are to be useful in bringing multiple judges to the same frame of reference. Selected exemplars for each of the seven scored tasks discussed in this book are included in this chapter.

A scoring session was originally scheduled for June, 1993, immediately following the testing. 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, such as 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 rubrics also needed significant revisions because some of the student work could not be scored.

In preparation for this meeting, each team was responsible for identifying 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[1] 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 were examined empirically and sparked insights into problems or successes in interrater reliability (e.g., Vermont Study, Rand, 1992) when the scoring was conducted. 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 revised workplan for the June meeting, the project staff brought in external science and mathematics educators who had not been active partners in this project to serve as consultants. 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. Each team 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

[1]These exemplars are available upon request from the Educational Testing Service; direct requests to M. Jorgensen.

were field tested with live papers supplied by the schools and, in particular, those selected as sample papers.

In general, the final changes to the rubrics include:

- Rewording to eliminate ambiguous language.
- Rewording to eliminate overlap between score points.
- Eliminating constructs that were no longer included in the task.
- 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 rubrics were applied to the sample papers originally selected by the teachers/developers. Whether or not the sample papers were scored at the same point or score location along the scoring guide was critical to the revision process because that was one way to sustain the intent of the original development effort. And, for example, the most able student response based on the original rubric continued to be the most able response using the revised rubric and likewise down the scoring scale.

With the revised rubric and re-selected exemplars, the project was ready to begin scoring of the student products in October, 1993. 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. 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 each rubric if 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 revisions, some major, 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 upon 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.

The sample papers presented in the following pages were selected to illustrate variation along the scoring continuum for each portfolio entry. Some of these were used as training papers but not all. And, many of the training papers are not included here for the sake of brevity.

Scoring

The training materials for the scoring were compiled and two teachers from each of the six school 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 people who were not familiar with the tasks and the development process.

Training began by reviewing each rubric one at a time and then 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 score assigned for the second and third sets of papers were not given and the readers were asked to independently score each paper. The scores were recorded and discrepancies (relative to the "true score" estimates provided by the development 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. The criteria for sufficient training was when the raters reported exact agreement 95% of the time, and those not in agreement were no more than one score point from the target.

Batches of approximately ten papers were prepared for the scoring session. For those portfolio entries with prescribed, and therefore common, content (*Retelling* and *Toys in Space*), the batches were built by randomly selecting student responses. This was done to reduce the bias potential when packets contain student work from a single classroom.

For the other portfolio entries that did have a different content stimuli, it was not possible to randomly select student responses across the classrooms. However, we did make every effort to group student responses so each batch contained a representative sample of the entire student group from each classroom. The practical realities of including portfolio entries that support teacher-selected content stimulus materials is that a description of that content must accompany the scoring packet and must be addressed uniquely in the training. These factors act against cross-group randomization. And, because the developers themselves were both the teachers and the scorers, no teacher was permitted to score student responses generated in their own classrooms.

Five of the tasks were scored during one full day and one-half day. The remaining three were scored off-site on two additional days. All papers were scored twice.

Based on the results of the scoring, final revisions to the assessments were made so that those contained in Chapter 4 are the end result of an iterative process of development, review, and revision. These final forms were the ones field tested in January 1994. The student responses during this larger field test held in 1994 provide the foundation for the discussion of interrater reliability, and performance transfer.

The project produced eight assessments. These were field-tested in the spring of 1994. Every effort was made to ensure that some students in each of the twenty-four classes taught by our project participants had an opportunity to perform on each assessment. The number of student responses by assessment is indicated in Table 13.1.

Table 13.1 Student Responses by Assessment Task and School

	Student Responses (N=1663)							
System	Science Observation	Retelling	Letter Writing	Continuum of Progress Towards Goals	Toys in Space	Problem Solving	Interview	Comparison of Experiments
Dade	15	25	13	11	14	19	5	0
Clarke	25	19	13	13	11	14	14	19
Marietta	63	56	56	32	28	55	31	22
Gwinnett	66	68	51	68	28	47	0	116
Richmond	64	70	92	49	41	28	0	21
Fulton	54	54	21	15	4	5	3	26
Total	287	292	246	188	225	168	53	204

Interrater Reliability

There are two types of reliability issues associated with performance task and, in this case, portfolio entries. One is the stability of the judgement made about the quality of the response. A second is the stability of the estimate of the individual's or group's responses (if aggregated). Without stability in the judgement made, there can be no stability in the estimate of individual or group stability. Thus, the reliability or stability/consistency of the judgment is a foundation without which no engineer or architect can craft a structure of individual or aggregate learning. But, on the other hand, establishing judgment stability does not ensure that the estimate of an

individual's performance is stable. But, beginning with the foundation, this chapter presents the data relevant to judge stability—often called interrater reliability.

Just as assessments must be systematic in their administration and in their documentation of evidence to meet the requirements of the paradigm underlying the AAMU project, so too must the judgements about the student products be systematic and comparable. That is the key rationale for developing scoring guides or rubrics that can be used by independent trained raters to make comparable judgements about the quality of student work. A common index of whether or not different trained raters do, in fact, agree one with the other about the quality of student work is *percent exact agreement.* In simple terms, this index asks the question "what percent of ratings from two independent judges are the same for samples of student work?" For the AAMU project, if there were 100 students responding to a task producing a product to be judged and if two raters produced the same score (index of quality) for 80 of those students, the percent exact agreement would be 80.

Percent agreement is one way to communicate the extent to which independent judges agree with respect to classifying student work within a context of quality as defined by a scoring guide. The National Assessment of Educational Progress (NAEP) employs this technique among others to indicate score stability. The higher the percent, the more often the raters agree. "The percent of exact agreement does not take into account the possibility that two raters might assign the same rating to a paper purely by chance, however."[2]

If you look at Table 13.2 and scan the column "Exact Agreement," you will notice that *Retelling* and the prediction score from *Toys In Space* stand out from the rest of the assessments with exact agreements of 76.9% and 74.4% respectively. These percent exact agreements are comparable with those from the direct assessment of writing.

[2]Kaplan, B.A. & Johnson, E.G. (1992, April 24) *Reliability of Professionally Scored Data NAEP-Related Issues.* American Educational Research Association, San Francisco, CA.

Table 13.2 Interrater Agreement Estimates

Assessments	Exact Agreement	Within One Score Point (Adjacent Agreement)
Letter Writing	42.2%	87.4%
Science Observation	46.0%	89.9%
Problem Solving		
Understands Problem	64.2%	96.3%
Plans/Reports Solution	61.1%	97.9%
Analyzes Results	62.6%	96.8%
Comparison of Experiments		
Understands Concepts	52.9%	97.1%
Extends Learning	70.0%	97.1%
Communicates	57.1%	95.7%
Continuum of Progress Towards Goals		
Focus	42.8%	83.4%
Strategies	45.5%	76.6%
Summarizes	60.7%	81.4%
Applies	66.2%	94.5%
Retelling	76.9%	99.0%
Toys in Space		
Prediction	74.4%	97.2%
Drawing	68.4%	88.4%
Narrative	50.4%	86.0%
Contrast	64.0%	92.8%
Question	57.6%	89.2%
Interview	NA	NA

Retelling and *Toys in Space* are the two portfolio assessment entries that had prescribed content. Thus, one might speculate that it is more difficult to train raters to make comparable judgments if content varies. This appears to be a finding similar to that discussed by Koretz and others (1992) in their Interim Report. For example, they report rather similar percent agreements for mathematics at both grades 4 and 8 across the multiple scoring criteria (see Table 13.3).

Table 13.3 Mathematics Composite Scores: Percent of Students for Whom Raters Assigned the Same Score[3]

Scoring Criterion	Grade 4	Grade 8
Language of Mathematics	52	49
Math Representations	55	56
Presentation	54	51
Understanding of Task	75	76
How: Procedures	66	62
Why: Decisions	57	49
What: Outcomes	81	89
Average	61	62

Although worded differently, these data represent the same index of stability—the information as the percent of students for whom raters assigned the same score and the percent of raters with exact agreement. So, relative to this index, the portfolio entries developed for the AAMU project elicited relatively the same challenges for raters. One can speculate on why independently trained judges did not agree as often on the non-content specific tasks such as *Letter Writing*, *Science Observation*, and so forth. It is likely that, as Koretz et. al., suggest in their report, that the training task becomes considerably more difficult when the content vehicle itself varies from classroom to classroom. In short, perhaps training materials were not illustrative of the kinds of decision challenges that faced the scorers. Another plausible explanation is that the variation in content made the tasks more or less appropriate and that the structure of the tasks themselves interferes with the ability of trained raters to judge the student responses comparably.

Unspecified content also appears to be a culprit—raters did not know what content was presented in sufficient detail to avoid having to literally guess about accuracy, completeness, and so on. Such "guessing" on the part of the judges may have also contributed to some scoring disparity. And, when we examine the index of scoring stability from the two assessments with content literally fixed in the AAMU project (*Retelling* and *Toys in Space*), the ability of judges to agree stands out dramatically as virtually an easier task. Keep in mind that the individuals doing the scoring were comparable with respect to their science background, familiarity with the tasks, and lack of experienced as scorers, and that no teacher scored student work from their own school. In this way, the objectivity of the judges was supported.

An extension of percent exact agreement is percent adjacent agreement (see Table 13.2). Indeed, in many projects where student work is judged by trained raters (e.g., state assessment programs for the direct assessment of writing), the index of credibility for the assessment program includes those in exact agreement and ratings of plus or minus one score point. So, if one rater judges the quality of work to be a "3" for example and another rater judges the quality of work to be a "4" for all intents and purposes, these are considered comparable judgments. We can argue the wisdom of this given score ranges of only four and five points but there are numerous examples in large-scale testing where comparable judgment is defined as a one-point difference.

[3]Excerpted from Interim Report (Table 7; same title)

A review of Table 13.2 reveals that for all but one scoring decision (*Continuum of Progress Towards Goals: Strategies*), the percent of exact plus adjacent agreement among raters exceeds 80%. This large shift from many raters not being able to agree exactly on the quality of stude work to most raters being able to agree prompts several questions:

- Are the category descriptions or training materials sufficiently clear so as to present clear and unambiguous examples of each of the levels of work as defined by the scoring guides?
- Second, are the categories described by the a score sufficiently independent from one another?
- Third, is the range of responses sufficiently restricted in variation so that the internal frame of reference each rater was using literally slipped so that some were prompted to "give the benefit of the doubt"—perhaps a one-point advantage because there were no clear-cut examples of higher levels of quality work?

Each of these questions remain unanswered from the perspective of this project but will likely resurface in any portfolio entry or performance-based assessment program where new tasks and new scoring guides are being developed.

Another typical index of interrater reliability is "reliability coefficient," a measure of the extent to which two raters rank students' work the same. It ranges from 0.00 (essentially, no agreement beyond chance) to 1.00 (perfect agreement).[4]

Selecting the correct correlation method depends upon the character of the data. Specifically, if the scores are interval (numbers with equal distances between them as in Normal Curve Equivalents (NCEs), then Pearson Product-Moment correlation is appropriate. If, however, the scores described in the scoring guide are ordinal in character (suggesting only more or less of the underlying trait such as 'predicting'), then Spearman Rho is appropriate. In reality, the decision about the character of the data is often a judgment call in the early stages of assessment development. Few developers have the resources or data sets to support scaling analyses. And, with respect to the AAMU portfolio entries, there appears to be a basis for arguing that some scoring guides are more interval-like and others more ordinal-like. For example, one could argue that holistic scores are more ordinal in character than interval. Conversely, one could argue that analytic scoring systems are more interval-like than ordinal. However, because no scaling analyses have been completed and because the real impact of the decision regarding which correlation technique is the better one, we have taken the position of offering both (see Table 13.4).

[4]Koretz, et al., 1992, pp. 2-3

Table 13.4 Interrater Reliability Coefficients

Portfolio Entries	Pearson Product-Moment Correlation Coefficient	Spearman Rho Correlation Coefficient
Letter Writing	.59	.54
Science Observation	.63	.58
Problem Solving		
Understands Problem	.47	.41
Plans/Reports Solution	.64	.62
Analyzes Results	.66	.66
Comparison of Experiments		
Understands Concepts	.68	.67
Extends Learning	.65	.61
Communicates	.55	.53
Continuum of Progress Towards Goals		
Focus	.55	.44
Strategies	.70	.69
Summarizes	.79	.80
Applies	.58	.57
Retelling	.86	.86
Toys in Space		
Prediction	.79	.80
Drawing	.62	.62
Narrative	.64	.60
Contrast	.78	.68
Interview	NA	NA

Again, it is useful to have some basis for comparison of these indices separate from, or outside the scope of, this portfolio study. Reliability coefficients of .70 or higher are not unusual for standardized performance assessments in writing (that is, assessments such as the Vermont Uniform Test of Writing, in which all students write responses to the same prompts.) Although reported as average reliability[5] coefficients across all seven scoring criteria in mathematics, the Vermont portfolio project provides one such frame of reference (see Table 13.5).

[5] "Because the scales used in rating the portfolios were only ordinal, Spearman correlations are reported throughout this memorandum. However, the more conventional Pearson correlations were only trivially different. (Koretz, 1992, p. 3 footnote 1.)

Table 13.5 Reliability Coefficients, Mathematics Composite Scores[6]

Scoring Criterion	Grade 4	Grade 8
Language of Math	.23	.28
Math Representations	.33	.31
Presentation	.45	.42
Understanding of Task	.26	.35
How: Procedures	.44	.30
Why: Decisions	.40	.31
What: Outcomes	.23	.35
Average	.33	.33

It is important to note that these values reported in Table 13.5 are based on composite scores rather than individual decisions such as reported for the AAMU project in Table 13.4. "Rater reliability would have improved modestly if students' overall score had been a simple average of their scores on each criterion, rather than the composite formed by the mathematics committee's algorithm. In fourth grade, the average rater reliability (across all criteria) would have been .44 rather than .33 if student's overall scores had been a simple average of their scores on all five pieces."[7] However, the bottom line for developers of portfolio entries is that we have a substantial challenge to demonstrate to our customers that the portfolio entries themselves and the portfolio as a meaningful and purposeful collection of evidence is credible and that the scores generated are equally as credible.

The distribution of scores across the continuum defined by each scoring guide also influences the meaningfulness and appropriateness of the correlation coefficient as an index of scorer consistency. The concentration of scores tends to depress reliability coefficients, even when raters agree much of the time. This phenomenon was also noted in the Vermont portfolio study.[8]
Table 13.6 presents the score distributions for *Letter Writing* and *Science Observation*. These are presented both for the separate grade levels and for the group overall. Equally normal distributions of scores were reported for the other tasks thus suggesting that the problem with rater agreement was not in compressed distributions.

[6] Excerpted from Koretz, et. al, 1992, p. 11, Table 5.

[7] Koretz et. al., pp. 11-12.

[8] Koretz et al., 1992, p. 13

Table 13.6 Score Distributions

Score Label	Value	Grade 3	Grade 4	Grade 5	Overall
Letter Writing					
No Attempt Made	0	13	17	8	38
Minimal Understanding	1	17	19	8	44
Limited Understanding	2	39	43	48	130
Satisfactory Understanding	3	43	74	68	185
Good Understanding	4	11	30	38	79
Exceptional Understanding	5	1	6	9	16
Science Observation					
No Response	0	3	42	8	53
Poor	1	11	29	32	72
Fair	2	64	76	49	184
Good	3	49	63	80	192
Very Good	4	22	16	22	60
Excellent Understanding	5	5	0	3	8

Clearly the direction in which to go to achieve credibility with respect to the stability of judgements made about the quality of student work is to develop tasks that are linked explicitly to specific content or to specific kinds of content like *Retelling* or *Toys in Space*. What is not known is whether the analytic scoring guides for each of these assessment entries provides the "key" to comparable judgments or whether the constancy of the content across groups of students provides the "key." These questions are easily researched. For example, one could replace "John o'Groats" with another stimulus—prose, visual representation, numerical expressions—that tell a story, conduct the scoring and add to the evidence regarding whether the scoring rubric is the key or not. Likewise, with *Toys in Space*, one could replace the NASA video with other equally rich science learning videos. Or, one could use a direction-conducted experiment as the stimulus.

Figuring out what the problem(s) is(are) on the other portfolio entries is not as straightforward. One approach might be to transform some of the existing holistic scoring guides into analytic guides, thereby focusing the decisions required on the part of the rater. Of course, in some instances the task does not lend itself to an analytic scoring rubric. The pattern of interrater agreement reported in Table 13.2 suggests that decisions on separate elements are more comparable than those generated on holistic scoring guides.

Regardless of the appraoch to scoring selected, i.e., whether analytic or holistic, the selection of content must be carefully made. In assessments where classroom teachers have the option to choose appropriate content, it is critical that they think through the match between what students have learned and what information is necessary to move towards a solution on the portfolio entry.

14

Value Added and Lessons Learned: People, Ideas, and Dollars

An important question to answer with respect to entries in portfolio assessment systems or any new forms of assessment is whether the behaviors documented generalize beyond the specific context of a single task. For the same reasons that new forms of assessment appeal to educators, they have the potential of being perceived as so novel, so unique, and so compelling because of their format, context, and presentation that the underlying skills of interest may be evoked only in those particularly limited circumstances.

A second perspective brought to this discussion is how do the points of evidence portrayed through the structured core of the portfolio paradigm augment, supplement, and enhance the existing picture of student learning?

The costs of researching and refining portfolio assessment in mathematics and science for this project took many forms. Along the journey, one individual resigned from the project. Others simply could not find the time to meet or to work as groups. Other costs included the creativity deficits experienced during development and refining of the tasks. Finally, the ultimate "real" cost of dollars for delivery of assessments for the structured core of the portfolio were dramatic. The question that remains is whether the process can be streamlined, the participants retained at an almost perfect level, and whether the payoff in terms of teaching and learning are significantly greater because of the assessment paradigm than would be gleaned through traditional assessment strategies. For this project, the summary remarks of the participants is a resounding "yes" to each of these questions.

Portfolio Entries as a Complete Picture of Learning

The spirit of portfolio assessment as advocated by supporters of whole language is that the purposeful collection of work will provide a complete and comprehensive picture of student growth in a discipline. Thus, the portfolio should contain benchmarks of beginning pieces, emerging or growing pieces and final versions of work—reflecting each student's understanding of finished, high quality work. Thus, as a collection of work composed of samples taken along each student's learning journey, a portfolio has the potential to be both comprehensive and complete. However, as discussed in Chapter 1, unless the samples are comparable from student to student there is little basis for aggregate analysis or for relative statements of growth or progress.

Our solution was to build a portfolio with two types of entries in it. First would be the core of the portfolio—indeed the assessment portion—that would be structured in ways to elicit data that could be aggregated or analyzed from some perspective greater than or broader than that of an individual student. For this discussion, this structured core was characterized by eight quite distinct and different tasks. Then, surrounding this core is the second type of entries, those that are descriptive of individual students and their unique learning journeys and that are the legitimate basis for individual descriptions of growth, progress, and concerns. This type may include such idiosyncratic entries as "best piece," "favorite piece," "most challenging piece," or even reflections and letters to portfolio readers as to why certain items were included by the student in the portfolio. None of these entries are likely to have the attributes that this author considers essential for them to be considered as or treated as assessment entries.

This approach of blending a structured core with the idiosyncratic selections of students and teachers is an extension of the Kentucky[1] model:

	On Demand	Extended
Uniform		
Local Option		

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 and other pieces of student work for which the line of evidence is less than direct.

The distinction between artifacts or evidence that has a clearly established line of evidence from the student to the work and those for which that relationship may be clouded -- as by an overly involved parent to siblings who have already completed similar work to students with resources to literally 'hire out' the work, perhaps best describes the press to include an uncontaminated structured core into the concept of portfolios.

It is this same logic that reminds us that 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

[1] *1991-92 Technical report*, Kentucky Department of Education, 1993.

also important to use the big ideas underlying reform as clarifying variables to enhance the process of schooling and, in turn, of assessment.

So, having presented and argued for a two-level portfolio system in which some entries form the assessment core and some form the individual descriptive core, it is important to advance the question of whether or not the entire portfolio as a purposeful collection provides a meaningful description of student performance and learning. Does the collection as a whole provide more than the insights gained by collecting the individual pieces? In other words, can we accomplish a higher-order description of learning by controlling, managing, or structuring all of the portfolio entries in such a way as to elicit a picture of the student performing in the discipline not possible in any other way?

This question caused the teachers and project partners in the AAMU to develop entries that were different in some respects while similar or common in others. Through this strategy it was anticipated that the image that would emanate would be of a student doing the work of science and of a student doing the work of mathematics. Some of the positive but moderate to low correlations reported in Table 17 suggest that this strategy has worked. Students do perform in somewhat comparable patterns from one task to the other but no task is so similar in the evidence it elicits that they are comparable tasks.

Assessment Tasks as Unique Stimuli

In the process of determining the nature or character of the portfolio assessment entries, the development teams attempted to vary the stimulus complexity, response complexity, content coverage, and cognitive complexity. The extent to which this goal was achieved is supported in the empirical evidence (see Table 14.1). For example, the low to moderate correlation coefficients reported throughout the matrix suggests that the assessment entries are capturing different kinds of evidence one from another. Certainly the performance of students on *Letter Writing* has little relationship to the performance of these same students on the *Continuum of Progress*. Similarly, *Problem Solving* and *Retelling* (.10) have little in common. In contrast, *Letter Writing* and *Problem Solving* and *Letter Writing* and *Comparison of Experiments* are related as are the *Continuum of Progress* and the *Comparison of Experiments* (.36).

This empirical information provides a platform for more questions. For example, one must wonder what it is about *Science Observation* that that particular portfolio entry elicits from students similar levels of performance as evidenced in each of the other tasks. In contrast, each of the other assessment entries has a less consistent relationship with the other assessment entries. Shared variance speaks to the connections across the work performed by this group of students. The fact that the positive significant relationships are low-to-moderate rather than high suggests that these assessment entries do perhaps elicit different kinds of evidence according to which task is the prompt. If this is indeed the case, then perhaps we have accomplished the goal of structuring portfolio assessment that examines the multiple facets or dimensions of science and mathematics learning through the use of different formats and structures for the entries. There is certainly enough evidence of this phenomena to justify continued study.

Table 14.1 Relationships Among Tasks

	Comparison of Experiments	Continuum of Progress Towards Goals	Letter Writing	Problem Solving	Retelling	Science Observation
Comparison of Experiments	1.00					
Continuum of Progress Towards Goals	.36 (37) p=.015	1.00				
Letter Writing	.42 (65) p=.000	.02 (115) p=.412	1.00			
Problem Solving	.40 (29) p=.015	.07 (55) p=.318	.49 (97) p=.000	1.00		
Retelling	.54 (54) p=.000	.30 (162) p=.000	.30 (172) p=.000	.10 (114) p=.155	1.00	
Science Observation	.37 (60) p=.002	.25 (189) p=.000	.25 (189) p=.000	.36 (120) p=.000	.37 (229) p=.000	1.00
Toys in Space	.37 (63) p=.001	.04 (130) p=.306	.26 (185) p=.000	.22 (111) p=.009	.34 (174) p=.000	.39 (209) p=.000

Costs and Benefits

For this project, the documented hours for which the twenty-four participating teachers were compensated exceeded 3500. In addition to these hours, all of the teachers contributed personal time and energy to ensure that the right portfolio entries were the best they could be. Added to this would be the project staff time generally estimated as 20% of two full-time equivalents.

If 3500 hours is taken as a minimum estimate of time invested in the development, field-testing, and scoring of eight tasks, that suggests that each task required over four hundred hours. Since each team responsible for a task had four members, this reduces to approximately 100 hours investment per individual contributor. However, at least half of this time was spent on professional development to support the work in assessment development. Subtracting the professional development hours, the time to completion of each task is approximately 200 hours total or 50 hours per team member. Thus, in full-time equivalent terms, each task or portfolio entry required, on the average, five weeks of effort. Added to that would be the costs associated with printing, distribution, and return of testing materials.

This project supported the teachers in two ways: through payment of substitutes when project work required that the participating teachers be out of the classroom, and through direct payment as consultants. During non-contract time, the teachers were compensated as consultants at a rate of $150.00 per day. If five weeks per portfolio entry is a reasonable estimate of time for development, refinement, and scoring, and if $150 per day is the typical cost of development talent, then the five weeks times $150 per day yields a per portfolio entry of $3750. When the professional staff time is added, the total estimate is $7595.

These costs, when compared with those reported by Hardy[2] seem reasonable, especially given the amateur status of the teachers with respect to their test/task development capabilities. Hardy offers some cost comparisons that help put this project's estimated cost for development, administration, and scoring in perspective. He cites cost estimates from the Kentucky proposal of $5,500 per exercise in the first year of work to $6,294 over the full five years of the contract, but there are many variables that prevent the direct comparison.

As the Kentucky statewide assessment program moves into its second contract period and as it completes its fifth year, it would be very interesting to find out the accurate costs associated with the development of performance tasks. However, it is reasonable to suggest that the teachers participating in this portfolio project were not far off target relative to effort required to complete task design and scoring as compared with one large-scale testing program. Of course, there are many unaccounted for variables in this analysis and we do not pretend to compare the quality of the portfolio entries with the performance events developed for Kentucky.

Certainly there are likely to be differences in process and in purpose. From either perspective, this project or the example of the larger statewide assessment program, it is not an understatement, however, to say that this work, like any good test development, is not going to be accomplished without substantial commitment of resources.

The true positive benefit of this development work is the change that happens in the developers. Whether we are talking about participating in the scoring of student work or in the entire development to administration to scoring process, the participants in this process change.

For this portfolio project, these participants were challenged to recognize that schools and school systems serving very different kinds of learners, staffed with very different kinds of teachers and administrators, could find common values for what students should be expected to know, understand, and be able to do in science and in mathematics. This common statement of explicit values through the project goals (see page 18), caused the teachers to recognize that the differences they had focused on at the local level were less important than the values they held in

[2]Hardy, R. A., Examining the costs of performance assessment, *Applied Measurement in Education*, 8(2), 121-134.

common for all students. This is not a trivial finding. In fact, it may be the single most important result of this portfolio project.

This is perhaps best said by a teacher:

> *The project has had a positive influence on classroom instruction and planning. We have a stronger understanding of how students learn best and how they can demonstrate that learning in a personal and realistic manner, and we were able to transfer this knowledge to teaching....This project has motivated us to communicate to students and parents the real life applications for multiple science and mathematics concepts....Participation in the project's activities assisted with the development of students', parents', and teachers' understanding that mathematics and science activities require an investigative and process approach to problem solving and experiments....We felt that one of the most important (aspects of the project) was the professional growth we experienced over the last two years. Having the opportunity to hear from pioneers in the field of assessment, interacting with science and mathematics experts, and making contact with other teachers around the state have all contributed to our growth as educators.*
>
> ...3rd grade teacher

And then the comments from a system-level administrator:

> *You cannot believe the difference in the classrooms of the three teachers from (system). My teachers did not want to teach science. They did not want to use manipulatives, nor did they see the value in using hands-on activities to make science come alive for their students. Now, because of their work on this project and because they have had the opportunity to read Science For All Americans (1989, 1990) and the NSTA preliminary standards materials and to talk with other teachers about how these habits of mind can and should be documented so as to provide unambiguous and credible evidence of the "big ideas" of science, these teachers have completely transformed their classrooms into environments where their students are "doing what scientists do"[3] in the best sense of that phrase.*

It also became clear through the course of this project that the findings reported by other change models proved to hold here as well. Specifically, key features for change include active participation, face-to-face interactions, opportunities to learn new behaviors, local materials development, and leader support.[4]

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 teacher views 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.

---

[3]Doris, E. (1991) *Doing what scientists do: children learn to investigate their world.* New York: Heineman.

[4]Osborne, B. "Creating a motivational learning environment," Paper presented at the third Annual Meeting of the National Conference on Creating the Quality School, Oklahoma City, OK, March 31-April 2, 1993.

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 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, 11*(1), 36-44.

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

Cronbach, L.J. (1970). *Essentials of psychological testing.* New York: Harper & Row.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 85*(4), 467.

Doris, E. (1991). *Doing what scientists do—Children learn to investigate their world.* Portsmouth, NH: Heinemann.

Eisner, E. (1992). The misunderstood role of the arts in human development. *Phi Delta Kappan, 73*(8), 591-595.

Hardy, R. A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education*, 8(2), 121-134.

Kaplan, B.A., & Johnson, E.G. (1992, April 24). *Reliability of professionally scored data NAEP-related issues.* A paper presented at the annual conference of the American Educational Research Association, San Francisco, CA.

Kentucky Department of Education. (1993). *Kentucky instructional results information system.* Dover, NH: Advanced Systems for Measurement and Evaluation.

Koretz, D., McCaffrey, Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program Interim Report.* Washington, DC: RAND Institute on Education and Training.

Moss, P. A., Beck, J. S., Ebbs, C., Matson, B., Muchlore, J., Steele, D., & Taylor, C. (1992, Fall) Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 2*(3), 12-21.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Osborne, B. (1993, March). *Creating a motivational learning environment*, Paper presented at the third annual meeting of the National Conference on Creating the Quality School, Oklahoma City, OK.

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-reflection. Unpublished manuscript.

Paulson, F. L., Paulson, P. R., & Meyer, C.A. (1991, February). What makes a portfolio a portfolio? *Educational Leadership*, 48(5), 60-63.

Porter, A.C., Archbald, D. A., & Tyree, A. K. (1991). Reforming the curriculum: Will empowerment policies replace control? In S.H. Fuhrman & B.Malen (Eds.), *The politics of curriculum and testing* (pp. 11-36). London: The Falmer Press.

Rutherford, F. J., & Ahlgren, A. (1989, 1990). *Science for all Americans*. New York: Oxford University Press.

Snow, R. (1989). Abilities, motivation, and methodology. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *The Minnesota Symposium on Learning and Individual Differences* (pp. 435-474). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stake, R.E. (1967). *The countenance of educational evaluation*. Springfield, IL: Gifted Children Section, Department of Exceptional Children.

Margaret Jorgensen received her Ph.D. in measurement, evaluation, and statistical analysis from the University of Chicago and has worked in the field of assessment in both theoretical and applied areas for over twenty years. She is currently a Senior Examiner in the Southern Field Office of Educational Testing Service.