DOCUMENT RESUME

ED 364 576 TM 020 797

AUTHOR Farr, Roger; Green, Beth
TITLE Indiana Performance Assessments '92. Final Report.
INSTITUTION Indiana Univ., Bloomington. Center for Reading and Language Studies.
SPONS AGENCY Indiana State Commission for Higher Education, Indianapolis.; Indiana State Dept. of Employment and Training Services, Indianapolis.; Indiana Vocational Technical Coll., Indianapolis.; Lilly Endowment, Inc., Indianapolis, Ind.
PUB DATE Sep 93
NOTE 74p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Assessment; Field Tests; Graphs; High Schools; *High School Students; Holistic Evaluation; *Mathematics Achievement; *Reading Achievement; Scoring; State Surveys; Student Attitudes; *Student Evaluation; Teacher Attitudes; Validity; *Writing Achievement
IDENTIFIERS Authentic Assessment; Indiana; *Indiana Performance Assessments; *Performance Based Evaluation

ABSTRACT

In 1991 the Center for Reading and Language Studies at Indiana University (Bloomington) assembled educational experts to investigate the validity of performance-based assessments for determining the achievement of high school students in reading, writing, and mathematics. The Indiana Performance Assessments that materialized from this project were subsequently used with over 5,000 students. This report describes the development of the assessments, their field trials, scoring, reliability, findings and test results, and how teachers and students responded. Of the 10 holistic assessments developed, 6 integrated reading and writing, and 4 integrated mathematics and communications skills. Given pilot testing constraints, student performance on these assessments was acceptable, and reliable increases in performance were seen from grade 9 through grade 11, with a significant increase beyond high school. The project successfully assessed over 5,000 students while conducting a survey and demonstrated that performance assessment is feasible on a large scale. The generally positive reception of the assessments indicates that both students and educators understand, or at least sense, the limited perspectives of much standardized testing, and welcome assessment in an integrated and authentic way. Results also indicate that reliability of scoring by a carefully-trained cadre is acceptable and the process quite feasible. Thirty-nine charts illustrate the discussion. (SLD)

Indiana Performance Assessments '92

Final Report

A Development Project Sponsored by:
Department of Workforce Development
Commission on Vocational and Technical Education
Employment and Training Services
Indiana Vocational Technical College
Lilly Endowment
Indiana Commission for Higher Education

Conducted by:
The Center for Reading and Language Studies
Indiana University

Principal Investigators
Roger Farr
Beth Greene

Writers/Developers
Cheryl Gilliland
Walter Hill
Erika Hopper
Julie McGee
Michele Peers
Marilyn Rindfuss
Bruce Tone
Elizabeth Worden

September 1993

Indiana Performance Assessments '92

Final Report

A Development Project Sponsored by:

Department of Workforce Development

Commission on Vocational and Technical Education

Employment and Training Services

Indiana Vocational Technical College

Lilly Endowment

Indiana Commission for Higher Education

Conducted by:

The Center for Reading and Language Studies

Indiana University

Principal Investigators

Roger Farr

Beth Greene

Writers/Developers

Cheryl Gilliland

Walter Hill

Erika Hopper

Julie McGee

Michele Peers

Marilyn Rindfuss

Bruce Tone

Elizabeth Worden

September 1993

Table of Contents

Preface i

Setting the Context: *New forms of assessment can help determine if the Education 2000 goals are achieved* 1

The Assessments: *Development of the Indiana Performance Assessments '92* 11

Field Trials: *Indiana tryouts involved over 5,000 students and several hundred teachers across the state* 25

Scoring System: *Establishing consistent scoring for a wide range of individual responses* 28

Reliability: *How the reliability of the scoring system was determined* 34

Findings and Discussion: *How did students perform on these types of assessments?* 43

Reactions: *Teachers' and students' survey responses to the assessments* 55

Summary: *What can be concluded about the use of this type of performance assessment in Indiana?* 63

Appendices

Preface

The use of performance assessment is a growing phenomena in schools across the United States. The debates about its value, technical suitability, and feasibility are both important and appropriate. None of us should be blind defenders of the latest panacea, nor should we abandon a promising change merely because all of the problems and issues have not been resolved. This pilot study is an attempt to provide some of the needed information about how such assessments can be developed and implemented.

The criticisms of performance assessment are often raised by those who ask questions about its technical soundness. They use the traditional short-answer tests as the model against which performance assessment should be judged. Technical soundness is indeed an important aspect of any assessment. However, the analysis of validity and reliability issues related to short-answer tests may require different perspectives from those for performance assessments.

The supporters of performance assessment argue that such assessments are needed because they support good instruction. Proponents suggest that performance assessment may have a profound impact on instruction merely because of the way it looks. Teachers often respond to assessment by attempting to develop the behaviors that are described by the assessments used in a school district or state; and since performance assessment has students actually apply behaviors like reading, writing, and computing, it seems worth offering as the model. Indeed, the phrase *measurement driven instruction* has been discussed extensively in the education literature and seems to be a well accepted phenomenon.

Regardless of one's views about performance assessment, it is important to point out that performance assessment is neither new nor unique. Such assessment has a long history in education and is currently widely used in health professions, the commercial and service sectors, and amateur and professional sports. Indeed,

all states in the country include a performance assessment as a criteria for securing a license to drive an automobile on public thoroughfares.

While we strongly support the development and use of performance assessment, we do not mean to suggest that performance assessment, or assessment of any kind, is the magic bullet that will solve education's problems. Some legislators seem to imply that assessment can accomplish this as they pass one legislative bill after another authorizing a variety of assessments to meet their accountability needs. Rather, we believe that assessment is an important aspect of education and needs to be studied and improved as do all other aspects of education. There is no question that excellent instruction is possible without any formal assessment. Some would claim abandoning formal assessment would make excellent instruction more probable, but despite concerns that would ban all tests, we believe assessment has a proper place in education to help inform students, teachers, administrators, legislators, and the general public about the status of learning in our schools. This project has been an attempt to learn more about the use of performance assessment as it helps to inform all of these groups.

We are very grateful to the excellent research/development team that worked very hard to insure the success of this project. They include the writers who developed the prompts: Walter Hill, Julie McGee, and Marilyn Rindfuss. Elizabeth Worden designed and laid out the assessment booklets. Erika Hopper and Michele Peers managed many aspects of the project, including the training of scorers, identifying tryout sites, and managing the papers as they were distributed and returned. The overall project manager, who kept all aspects of the project moving along, was Cheryl Gilliland. Bruce Tone took on major responsibility for formatting and editing the final report.

We are especially pleased with the support and encouragement we received for this endeavor from our project sponsors: Clyde Ingle and Karen Rasmussen from the Commission for Higher Education; William Christopher and Peggy O'Malley from the Department of Workforce Development; Donald Warren, Dean of the IU School of Education; Kent McGuire from the Lilly Endowment; and especially Stan Jones and Nancy Cobb from the Governor's

Office. Their interest and enthusiasm for the improvement of education helped to maintain the momentum for the project. Without these people this project would not have been conducted.

We would also like to thank the many students, teachers, and administrators from the schools throughout the state and across the country who took time from busy schedules to review materials, participate in the tryouts, and help to assess responses.

The project was a pilot project to determine the feasibility of this type of performance assessment on a large-scale basis, and this report describes the results of that tryout.

Roger Farr
Beth Greene

Center for Reading and Language Studies
September 1993

Setting the Context: New forms of assessment can help determine if the Education 2000 goals are achieved

Many states and school districts across the United States are moving to performance-based assessment, which is purported to be a more authentic evaluation of both student achievement and school accountability. This movement has been promoted by influential national documents on education, beginning with the articulation of goals in education to be achieved by the year 2000 and followed with reports such as that from the Secretary of Labor's Commission on Achieving Necessary Skills (SCANS). The America 2000 goals, which were articulated by President George Bush and the nation's governors, called clearly for *performance assessment* to determine and report on progress in achieving them; and the SCANS report underlined the importance of performance to assure that educational training translates to the ability to apply skills needed in the national workforce.[1]

Early in 1991, the Center for Reading and Language Studies at Indiana University assembled a team of educational experts to investigate the viability of using performance-based assessments for determining the achievement of high school students in reading and writing and in mathematics. The assessments that materialized from this project and that were subsequently tried out with over 5,000 Indiana high school students[2] have been collectively named *Indiana Performance Assessments '92*.

---

[1] The goals are presented in *America 2000: An Education Strategy*, U.S. Department of Education, 1990. The Secretary's Commission translated the goals to workforce performance needs in "What Work Requires of Schools: A SCANS Report for America 2000," U.S. Department of Labor, June 1991.

[2] Approximately 20,000 students in Tucson, Arizona, and several school districts in Colorado and California were also included in the tryout although the results from those tryouts are not included in this report.

The overreaching questions of the study were:

- *How well can holistic assessment meet state-mandated needs? Would the holistic scoring that performance assessments use produce results that can reflect meaningfully on the quality of high school education available in Indiana?*

How well can holistic assessment meet state-mandated needs?

- *Can we create materials on topics of authentic interest and value to Indiana's secondary and post-secondary students which will assess their ability to think, read, and write, and to compute and apply mathematical concepts?*
- *Would such performance testing answer growing concern about the standardized and criterion-referenced testing now in prevalent use across the nation. Can it measure language use, applications of mathematical understanding, and thinking as the kind of processes described by cognitive theory that has developed over the past several decades?*
- *How would teachers and students react to a performance assessment which is structurally different from the standardized, multiple-choice tests that have been used in their schools and for the Indiana Statewide Testing for Educational Progress (ISTEP)?*

What led to the demands for more assessment?

School systems and society in general have always been interested in assessing the effectiveness of the education they support and provide youngsters. Since the development of standardized norm-referenced and criterion-referenced tests, most students being tested have responded to questions by picking the correct answer or completion of a statement from among multiple choices provided by the test makers.

Major crises have greatly intensified the general public's interest in educational assessment in the past several decades, including the launching of Sputnik by the Soviet Union in the early 1960s. Since that time, Americans have

been unusually conscious of how the academic performance of their students compares to that of other nations. In addition, a continuing decline in scores on college admission tests, in particular the *Scholastic Aptitude Tests,* was widely publicized; and the public became convinced that American students were learning less and less in schools that were costing more and more.

In response to this negative view of what students have been learning, most states established *competency testing*. It has been designed to identify students who need extra training and to hold schools, teachers, and students more accountable. Most of these tests are *criterion-referenced* instruments because teachers and others on committees make decisions about what *minimum essentials* are to be covered with multiple-choice items and what scores will determine competency.

In response...most states established competency testing.

Why the call for alternative assessments?

There has been a significant increase in the amount of testing and the time devoted to it in our schools. Many educators have been concerned that this has lead to a heavy emphasis on skill-drill type of instruction that requires the student to practice very short-answer responses to the skills and content tested. Classroom instruction frequently mimics the structure of the tests. Thus students get few opportunities to apply their thinking and language behaviors as ideas they had to construct themselves. In writing instruction, such an impact may be easily noted. In place of opportunities to actually write, some students are practicing usage skills isolated from any *meaningful* usage.

Embracing performance assessment does not imply that educators must throw out multiple-choice tests altogether; however, educators and the general public may be receptive to it because they are coming to recognize that testing for knowledge of highly specific subskills and specific subject material leads to

the narrowing of what students are taught. It is only natural that teachers, schools, and their curricula may emphasize what is on such tests that hold them accountable, seriously limiting the educational experience of our children.

This tendency has been linked to the criticism of standardized, multiple-choice tests which argues that they are *not* authentic measures of educational goals which aim at educating students to successfully apply what they learn. Few life tasks and experiences involve selecting the correct "canned" answer or solution from among several meant to "distract" the individual.

Few life tasks and experiences involve selecting the correct "canned" answer....

One of the most significant trends in instructional design and in curriculum in recent years involves *integrating* language and thinking behaviors as well as information and understanding that had previously been presented to students as separate disciplines in relatively isolated subject areas. This new emphasis emulates effective teachers who have students write about what they learn in science and apply mathematical concepts to ideas in social studies and the other humanities

These developments are consistent with established theories concerning the ways that language and thinking operate and develop. There is wide agreement among educators about the need for *authentic*—real— instructional experiences, and there has been an equal demand for assessments that are authentic.

The military services and industry have understood this argument for a long time, and the development of "hands-on" assessment has taken place in these arenas. The objectives outlined in *America 2000: An Education Strategy* are clearly expressed as performance goals. In discussing assessment, this call to action says, "Achievement tests must not simply measure minimum

competencies, but also higher levels of reading, writing, speaking, reasoning, and problem-solving skills." (p. 71) And this emphasis becomes dramatically clear in the report from the *Education Secretary's Commission on Achieving Necessary Skills* (SCANS). Its perspective on *Education 2000* translates its broader goals into precisely articulated *particular* performance tasks paramount to success in the workforce.

The Indiana Legislature has been alertly tuned to these developments. *Indiana House Bill 1695*, the Conference Committee Report on it, and *Senate Enrolled Act No. 419* use the term *performance* in describing assessment and the development of statewide assessments. They call for both a test of performance and the use of student portfolios, another performance assessment technique. Portfolio assessment is less formalized for educational accountability than is the more structured performance assessment developed and examined by this study, but it is equally valuable.

This legislative action has placed Indiana amid other states which are currently investigating, constructing, or experimenting with alternative performance assessments, most of which are intended to accompany state assessment programs already in place. These states are Arizona, California, Connecticut, Kentucky, Maryland, New York, and Vermont. Because of its experience in constructing such assessments, the Center for Reading and Language Studies became involved in an investigation of how performance assessment might serve the state.

...legislative action has placed Indiana amid other states which are currently investigating...alternative performance assessments....

What is performance assessment?

The experiences of other states indicate that performance assessment can be a viable alternative in Indiana's determination to better educate its citizens. Most states that are experimenting with and investigating performance assessment options are looking to it as a means of assessing major educational outcomes not covered by short answer or multiple-choice formats.

Performance assessment may take varied forms, ranging from observations of students conducting science experiments, scheduled stage performances, and exhibitions to the paper and pencil type developed in this project. The idea of performance assessment in the arts is a long-established tradition. Most citizens are familiar with the artist's portfolio and the musician's recital.

One popular type of performance assessment used in some schools consists of collections of student work (portfolios) which can be analyzed by the student and teacher on an ongoing, regular, and informal basis. While some states[3] are experimenting with portfolios for large-scale accountability reporting of school, teacher, and student performance, the structured type of performance assessment studied in this project appears to be more viable and acceptable to audiences interested in accountability because its structured system for rating student performance can be made reliable.

The *Indiana Performance Assessments '92* provide students with realistic (authentic) problems that ask them to read a variety of texts and to develop and write a thoughtful response to the problem for a particular audience. Students are encouraged to take notes while reading, to organize their notes, to develop a first draft, and then to revise and edit their response as a final draft. This performance assessment stresses the construction of meaning through reading

[3] For several years, Vermont has been engaged in a study of the use of portfolios for large-scale statewide assessment.

and writing as a thinking *process*, which is just what the use of language is now understood to be. Thus the task is designed as a realistic activity that approximates what students are expected to do in school—and in the world of work, as well. The student response is rated or scored on several key dimensions, using criteria that have been carefully developed and clearly articulated and referring to an extensive set of actual student responses that have been carefully selected as examples of all levels of student performance.

This performance assessment stresses the construction of meaning through reading and writing as a thinking process....

Many citizens interested in education who have reviewed this form of assessment seem to be supportive of the realistic focus for the assessments. Business people, in particular, have responded that they are far more interested in whether potential employees can perform these kinds of tasks than they are in whether students can identify right answers on multiple-choice tests. The administration of Indiana University[4] has also expressed strong interest in this type of assessment as a valid means of determining whether students can perform the kinds of literacy and thinking tasks expected of college students.

It is clear that literacy demands for both higher education and the workplace call upon students to use diverse thinking processes to accomplish a wide range of tasks. This usually means reading various information sources, identifying relevant information from those sources, and then using it in an organized manner.

The tasks presented on performance assessments can cut across subject matter areas and various types of texts, such as poems, tables and charts, essays, and letters. A variety of texts can be combined to depict ideas and problems in

[4] In the 1993-94 academic year, a pilot study of this type of performance assessment will be conducted with Indiana University undergraduates in the College of Arts and Sciences.

science, government, geography, and other subject areas. Writing the response can require that students combine a variety of thinking as comprehension-building while the information read is integrated into what the student may know about the topic.

Two examples developed at the Center illustrate briefly what performance assessment is:

- Middle school students are asked to read a sheriff's skiing accident report, to examine a map of the area where the accident took place, and to read a report based on an interview of one of the persons involved in the accident and what followed. The material is read in order to write an article for a junior high newspaper at the school where the boys involved in the accident study.
- Another example asks secondary school students to read an excerpt from Jules Verne's 1865 fictitious description of a trip to the moon and an excerpt from astronaut Michael Collins' report as one of the first men to actually travel to the moon. The students are asked to write a comparison of the two while predicting the future of space exploration.

This type of assessment appears to get at the behaviors that good instruction targets. Its use as a high school exit exam is not only a viable idea; it appears quite reasonable to argue that it is essential if we are to refocus educational systems on teaching students to *apply and demonstrate* their knowledge and skills. This study has investigated how viable that contention appears to be. A major goal was to begin developing probable answers to these kinds of questions:

This type of assessment appears to get at the behaviors that good instruction targets.

- What is the tradeoff between the broad, but shallow, look at student abilities provided by the short-answer and multiple-choice

test and the narrow but in-depth look that the kind of performance assessment being considered seems to provide?

- What kinds of reading texts do teachers, students, and those who want schools to be accountable consider to be authentic or real? In their demands for comprehension and their invitation for application, what topics that cross subject areas will be most real, meaningful, and engaging for students?
- What kind of writing tasks are the most authentic? What real audiences can be targeted?
- How can such assessments be administered? How much time should be allowed? Should students be allowed to consult the teacher, sources, and each other as they would on similar tasks during instruction?
- What format should be used for the presentation of the assessment?
- How much guidance should be provided to the student in prewriting and revision activities?
- What criteria should be used in rating student responses on the assessment? Which are compatible with the need for an authentic experience? How assessment-specific can or should they be?
- How can teachers best be trained to use them? Can the scoring be developed so that scores are reliable for individual student papers? How much training is needed to develop reliable scoring? How feasible will it be for local school districts to train their own raters and to score their own assessments?

While it was not expected that this pilot study would produce the *complete* or definitive answers for these questions, it lived up to expectations that it would give very valuable indications for developers of subsequent performance assessment that may be used in Indiana.

How did this study set out to answer these questions?

To answer these questions, the Center for Reading and Language Studies followed these basic steps:

- Ten assessments were developed and designed to cover the integration of reading and writing behavior, while including science and social studies, and to cover an integration of mathematics and communication skills.
- The assessments were administered in a broad sample of Indiana high schools and at a limited number of post-secondary sites to over 5,000 students.
- A scoring system for rating the student responses was developed to apply the rating range *1, 2, 3,* or *4* to each of several dimensions. Criteria for evaluating and assigning scores on each dimension were carefully articulated in rubrics; and the rubrics guided the selection of two *anchor* papers to exemplify each possible score for each dimension on each assessment.
- Paid volunteer teachers and school administrators were brought to the Center and trained to use the rubrics and anchors. Discussions about the rating task were held with volunteer raters.
- The project staff compiled, analyzed, and evaluated the information gathered during the project, including the scoring results and the survey responses, and noted indications and conclusions supported by the study.

The Assessments: Development of the *Indiana Performance Assessments '92*

The Center for Reading and Language Studies developed a total of ten holistic assessments. Six of these integrated the reading and writing behaviors of the subjects; the other four integrated math and communication skills. In conjunction with writing the assessments, the staff had four other important tasks to complete in order to prepare the assessments for the field trial:

- Prepare prewriting and post-writing (revision) activities as options for the students being assessed on the integration of reading and writing to follow in developing their responses.
- Develop a reasonable scoring range that could be assigned to the various aspects of the papers that were to be assessed.
- Identify dimensions widely acceptable to good teachers on which the student response would be scored, and consider generic factors that would define those dimensions in a table, or rubric, as a set of criteria for the evaluation.
- Prepare the assessment booklets and instructions for the students and the instructions to the test administrators.

Specific steps were followed in developing the Indiana Performance Assessments '92.

After consulting with content area education specialists and talking at length informally with numerous teachers, the project staff at the Center spent several weeks investigating, proposing, and discussing possible topics for the six reading-writing assessments and the four mathematics assessments.

Discussion at the Center coincided with careful consideration of how the assessment package should be structured. Many hours were spent discussing the viability of possible topics and the possible texts that could be used to

present them. Were the developing topics of interest to high school students? Did they suggest a realistic writing task? Did the texts that might be used support the task? As this selection and defining discussion progressed, the staff created a broad structure for the entire package.

Many hours were spent discussing the viability of possible topics and the possible texts that could be used to present them.

The aim of the Center staff was to insure that the reading-writing and math-communications assessments would be relatively compatible in format, style, and challenges to the students. The reading-writing assessment topics were selected first. The challenge was then to come up with mathematics assessments that were thematically related to those for reading-writing that were workforce and subject area related.

Because there were significant differences in the two general assessment categories, however, the descriptions of the reading-writing and of the mathematics-communication assessments are presented in separate subsections that follow in this report.

Six reading-writing assessments were developed.

The goal for the reading-writing assessments was to develop six that could be tried out in schools; and they were structured into three general genres:

- **Two required reading, analyzing, and responding to literature.** This type assesses those general reading processes which are needed in the reading of most prose. In these literature based assessments, the emphasis centered on the integrated reading-thinking-writing process needed for understanding and appreciating ideas presented in literary form and for extending them as an understanding of life. Students might be asked to compose journal entries, story endings, or critical reactions based on a synthesis of several text selections.

- **Two assessments required integrated responses to texts that could be involved in workplace activities and centered on using real-world literacy to perform tasks presented in workplace scenarios.** This assessment of practical literacy was designed to examine skills and strategies needed to obtain and succeed on a job and to advance there in an increasingly challenging and competitive job market. The performance tasks determine whether students can read, understand, and select relevant information from texts such as memos, reports, manuals, and graphic materials. They use their understanding of that information to write such things as sets of directions, job orders, interdepartmental memos, and basic summaries.
- **Two assessments relied on texts, comprehension, and application relevant to multiple content areas including science or the social studies.** This type was designed to assess reading skills and knowledge building that relies on contents in science, social studies, and other subject areas. The tasks involve reading such sources as textbook material, magazines, newspapers, history and biographical sketches, and/or tabular/graphic information. Students can be asked to write a report, a summary, an analysis, a letter to a decision maker, a recommendation for action, or some other type of writing.

As the reading-writing assessments were being developed, the individuals responsible for writing and designing them made regular presentations of what was materializing to the rest of the staff; and group reactions played a major role in the final selection and shaping of the six that were completed.

Student responses to the reading-writing tasks were designed as authentic uses of language.

In all six assessments, the students were asked to assume a situation that was designed to be realistic and plausible. A writing task that relies on comprehension of several texts was prescribed as a written response with a

particular audience, purpose, and format. The students were given specific guidelines as to how their work would be evaluated and scored.

The formats and ultimately the scoring criteria were designed to allow the student to demonstrate an ability to go beyond the basic requirement of the task. Analysis of the scores that were given by the teachers who scored the responses, and discussions with these teachers, indicated that a student's ability to go beyond the assigned tasks and to "take ownership," as some raters described it, was cited most often as meriting the highest grades.

As the topics and tasks for reading-writing assessments were structured, it became necessary to answer several key questions about how the assessments would be structured for the students:

- **How much writing assistance would be offered?**

As the assessments took shape, prewriting activities were developed to fit the context, situation/problem, and purpose of the writing task. Although these were to be optional, they were designed as relatively complete guides to comprehending, organizing, and reacting to the material in the prompts so that it could be used effectively and selectively in fulfilling the writing task. All of them presented note taking forms the student could use.

...prewriting activities....were designed as relatively complete guides to comprehending, organizing, and reacting to the material....

Questions were also developed to guide the student who elected to use them as a self-check after finishing a first draft of his or her response and before writing the final draft. These post-writing questions had to be revised and checked once the criteria for scoring the response were completed to be sure that they covered or directed the student to consider the key aspects or factors that described the dimensions in the scoring rubric.

While using this assistance was optional, it was available to structure the note taking and thinking the student might do about the texts read in the light of the writing task. These reminders for prewriting and revision were carefully discussed by the staff panel to help assure that they indeed did help the student and would not act as a hindrance to the writing process. One purpose of the study, however, was to determine if this was the actual case in a field trial.

- **How would the responses be scored?**

In order to provide a viable scale for discriminating among responses and at the same time to avoid too complex a scoring system, it was decided that a four-point scale (*1-4*) would be used for all of the Indiana assessments. With at least three dimensions, this would create twelve possible score descriptions for the rater to consider. Too complex a scoring system places unreasonable demands on those who must learn to score the responses and on the systems that train them.

- **What dimensions could help define a good response?**

In a task that would initiate and guide them in building the criteria for scoring the results, the project team consulted, discussed, and synthesized available research on holistic scoring to help identify the dimensions by which student responses would be evaluated. This effort was also based on experience in designing other integrated performance assessments and after talks with content area specialists and a number of teachers.

Three dimensions of the reading-writing assessments were selected. After considering various ways of factoring the evaluation of the assessment results, a simple and direct, but encompassing, system was devised. Student responses to the reading-writing assessments would be scored on:

- *Accomplishment of task,*
- *Reading,* and
- *Writing.*

(It should be noted here that there are four, not three, dimensions for scoring the math-communications assessments as described later in this report.)

- **How should all of this material be presented to the student?**

The reading-writing prompts and all the related materials were structured into booklets with directions to the students. The students were given a detailed statement as to how their work would be evaluated. For the reading-writing assessments, a general statement first explained the three dimensions as "skills" that would be evaluated and scored.

In reading, the student was reminded to understand the problem, to choose material read that was appropriate, and to interpret information from the text appropriately. In writing, the student was reminded to be clear and interesting, to vary sentence structure, to use correct conventions, and to relate parts of the writing clearly. The task stressed achieving a clear organization, developing solid main points and good supporting detail, meeting the needs of the audience, and using one's own relevant ideas to fulfill the task. This same list of evaluation factors was used on each of the six assessments.

A more specific set of directions to the student differed for each prompt. The students were given information about what they were to do under several headings. These instructions were a kind of "scenario" which explained the role the student as writer was to assume, the audience to be addressed, and the purpose for writing. A section of the instructions headed "Setting," explained the situation. Under the "Task," students were told the purpose and audience for writing and were given a list of three to five selections to be read and studied.

Under "Response Guidelines," the students got brief but more specific advice on what they should attempt to do, including a reminder to use standard mechanics. It was believed that the assessment should evaluate the student's

ability to comprehend and follow directions, so teachers were directed not to discuss with students the Setting, Task, or Response Guidelines.

As the section on results will show, both teachers and administrators praised the support built into the assessments; some teachers commented that the process was much like methodologies and materials they use in their classrooms. Many students also thought that the organizational keys were critical in helping them respond effectively. One student commented, *"This is the first time I ever learned anything during a test."*

...some teachers commented that the process was much like the methodologies and materials they used in their classrooms.

Finally instructions to those who would administer the assessment were prepared. These included hints and information relevant to successful administration and use of the results and numerous instructions for familiarizing students with the assessment packet and procedures, including the directions that the administrator should:

- Tell the students what is required of them and how their final product will be scored, assuring them that there are no right or wrong answers and calling attention to guidelines printed for the student in the packet.
- Explain what the students will be doing and emphasize that what is written should clearly reflect what is read.
- Allow three hours for the assessment, cueing students to begin their final draft, if they have not already done so, when there are only 30 minutes remaining.
- Allow the students to use or not use the prewriting and self-check activities.
- Allow the students to consult resources like dictionaries and to ask questions of the assessment administrator.

The instructions also covered any necessary instructions and guidelines relevant to specific assessments.

What topics and tasks were selected for the reading-writing assessments?

A brief description of each of the six reading-writing assessments indicates how they integrated reading, writing, and thinking.

Responding to Literature: The responses to the genre that used literary pieces as the reading prompts indicated that while the students found them to be more difficult than the other four, all students were able to express themselves, few were totally off task, and, to some extent, their responses were more individualistic than responses to the other genres.

- One literature-based assessment entitled *Introductory Essay on Relationships,* asked students to write the introduction to a section of a high school literary magazine. The section to be introduced included four short pieces: a poem about two people wanting the same parking place; a poem about social/political relationships; a section of a short story about a woman living in, and coping with, a changing neighborhood; and a short story about the relationship between two young boys.

  The students were clearly instructed that this section in the magazine dealt with personal and group relationships, and that their introduction should show how the pieces dealt with that general subject and belonged together despite their diverse subject matter.

- Another literature-based prompt entitled *Sports Essay* instructed the students to write an essay, as if for their English teachers, answering the question: *Is the emphasis placed on excelling in competitive sports healthy for American youth?*

  The texts were a poem about a high school athlete without marketable skills, a newspaper article about basketball coach John Thompson, and the introduction to a book on female tennis players.

Workplace Focus: Almost all students were able to incorporate clear evidence of having read the material in the two assessments designed to reflect real-world literacy tasks and to frame in-depth responses, but many failed to show much individualism.

- One of these assessments called *Purchase Recommendation Memorandum* asked students to respond to a memo from their boss, a store manager requesting advice on what kinds of bicycles to stock.

  The texts were the memo from the boss, a set of notes and tables taken from three issues of a consumer reporter magazine, and a business magazine article with information about different types of bicycles.

- Another of these assessments was called *Coordinator Hire Report* and asked students to pick a new employee as office coordinator from among four candidates and to write a memo to partners explaining the reasons for their choice.

  The texts were a company document explaining its needs and procedures for hiring, a document explaining some of the job responsibilities, the ad run to attract applicants, and the resumes and application letters from the four candidates.

Integrated Content Areas: The two assessments designed to integrate material from different subject areas framed actual problems existing in the United States. Students found them the most challenging of the six assessments, and their performance on these, therefore, indicated a broader range of abilities, especi. lly in *Accomplishment of Task* and *Writing.* These assessments also proved to be clear indicators of how well the reading materials were comprehended.

- One prompt, the *Water Problem Essay,* asked the students to write a letter to the California Water Commission suggesting a long-term solution to the state's water shortage.

The texts included a long encyclopedia entry on water, a news article on California's water problems, and a science news magazine article on technological solutions to water shortages.

- A second assessment in this genre, *Fuel Tax Letter to the Editor,* asked students to write to their local newspaper editor taking and supporting a position on whether a fossil fuel tax should be passed and if so, whether it should be modified from the major proposal.

  The texts included a newspaper article about the probable effects of such a tax, a textbook explanation of the environmental damage and other external costs associated with fossil fuels, a magazine article about "external costs" related to energy, and a brief governmental report about the probable effects of increased fuel prices.

Mathematics-communication assessments were based on a model.

Mathematics-communication prompt development was based upon the model known as *guided investigation.* As the term implies, an assessment of this type focuses on a central, relevant theme and uses questions and problems to lead students through a mathematical investigation of the thematic topic. The materials used in developing the prompts involved library and current event resources. All prompts were researched for accuracy. The topics for the math-communications assessments were chosen as "connectors" to the reading-writing portion of the assessment. The author of these assessments studied the reading-writing prompts very carefully and crafted mathematics investigations as complements to the four assessments in:

- real-world literacy and
- integrated content areas.

The response to the math-communication assessments was designed as problem-solving.

The math-communication assessment mirrors instruction: As students investigate the central theme, they learn about it; and they use the mathematics

skills that they possess to solve problems, to make discoveries and predictions, and to draw conclusions.

The situations that students encounter in the investigations are related to the real-world, and the work that they do during the investigation represents means to a very clear end. Students see that there is a reason for doing the investigation. In addition, the problems and questions are of the type that students as workers in the real-world are expected to solve; and the skills necessary to solve the problems are skills that all high school students are expected to possess.

...the problems and questions are the type that...workers in the real-world are expected to solve....

Under *Showing Your Work,* students are instructed to include all the calculations needed for each problem, writing down even the figures that might have been acquired using a calculator. They are told "to explain your reasoning in complete sentences" in order to get "credit for everything you do that is relevant...."

- **How much assistance would be offered?**

There are no pre-writing/problem solving activities in these assessments *per se;* however, the student is given helpful advice and graphic aids that are pertinent to solving the problems as they are presented. Some information pertinent to the particular assessment topic is presented with the "Directions for Students." For example, in the first assessment described below, there is a diagram of a racing bike with all its parts labeled, and a note instructs the student to use 3.1416 for pi. The assessments

..the student is given helpful advice and graphic aids that are pertinent to solving the problems....

include numerous graphics critically related to the computations and procedures solicited.

The students are advised that they should work the problems in order, can consult dictionaries, may use calculators, should read and reread, may skip through problems, and may go back and work on previous problems. They are allowed to ask the teacher for certain types of help, but the teacher is not to tell them if they are working the problem correctly.

- **How were the responses scored and on what dimensions?**

A four-point scale (*1-4*) was used for the math-communications prompts, but in scoring responses, *four dimensions,* rather than *three* as in the reading-writing assessments, were used. The four dimensions are:

- *Reasoning,*
- *Understanding of mathematical concepts,*
- *Communication,* and
- *Use of procedures.*

- **How were the math-communication assessments presented?**

Like the reading-writing assessments, those that covered mathematics-communication were printed in booklets that provided background about the particular assessment topic as *The Setting.* Then the students got instructions for *Working the Problems* that were the same for all four assessments. These provided directions about the assistance they could seek and use as described above. The instructions stressed the importance of showing their work and how they got their solutions to the problems.

The instructions stressed the importance of showing their work—how they got their solutions to the problems.

The math-communication topics and problems were related to those in the reading-writing assessments.

While the math-communication assessments were not paired to the reading-writing assessments in the field trial, each math-communication assessment is matched topically to one of the four reading-writing assessments in *Real-World Literacy* and *Integrated Subject Area* content. Conceivably the pairs could be given to the same students within a reasonable time frame to assess math-communications as well as reading and writing.

- One assessment, *Bicycle Racing,* explains the importance of gears in racing and gives descriptions, a list of terms, and a series of equations. This is followed by fifteen problems with room left to work and explain each. This assessment is a topic match for the *Purchase Recommendation Memorandum.*
- A second assessment, *Security One,* is a topic match for the *Coordinator Hire Report.* It describes a survey of potential customers run by a company that makes a home security system. The results of the survey are given in a table, and twelve problems follow.
- A third assessment, *Antarctic Icebergs,* is a topic match for the *Water Problem Essay.* It describes a proposal to tow icebergs from Antarctica to California to solve the state's water problems. The information to solve the problems, however, is presented with the problems themselves. This information includes a projection map, a note, and a conversion chart. There are thirteen problems.
- The fourth assessment, *Oil Spills, is* a topic match for *Fuel Tax Letter to the Editor.* It explains and presents three considerations about a proposed solution to a problem in Texas: the amount of oil dumped into storm drains by individuals who change the oil in their cars themselves. It also presents a series of problems related to the Exxon Valdez oil spill off the coast of Alaska.

The problems presented in each of these assessments were a sampling of the skills, problem-solving strategies, types of computations, and procedures that the *National Council of the Teachers of Mathematics*[5] has deemed as essential for high school students to acquire. These were built into the word problems and scenarios presented. The intention of the design was to present problems in a sequence that would not penalize students in later problems for errors in early responses.

In general, student comments about the mathematics-communication assessments were surprisingly negative across the four, as will be reported in a major section on Results. This general reaction appears to relate to the fact that the math-communication assessments required a considerable amount of highly focused and sustained thinking and work and because many students found them different from activities they were used to experiencing in math instruction. The student reactions, in turn, appear to have colored the reactions of the teachers who administered the tests.

---

[5] These essential skills are presented in National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: NCTM.

Field Trials: Indiana tryouts involved over 5,000 students and several hundred teachers across the state

The schools and students that participated in the field trial of the ten assessments were selected in an attempt to represent the diversity of the population of Indiana. Over 5,000 students from 28 high schools and several Ivy Tech campuses were included in the field trial. The sample was chosen to include students from grades 9 to 12 and post-secondary students, from a wide range of socioeconomic backgrounds, including a range of ability levels, and from a wide geographic distribution across the state.

As examples, some of the students attended suburban or rural high schools, such as those in Munster, Columbus, Goshen, and suburban Indianapolis; others were from urban schools, including schools in Fort Wayne, Indianapolis, South Bend, and New Albany. Most of the subjects were in the ninth, tenth, and eleventh grades. Students from vocational schools affiliated with the high schools in Indianapolis, New Albany, and Anderson were included; post-secondary subjects came from vocational schools located in Shelbyville, Bloomington, Sellersburg, and Indianapolis.

A small preliminary tryout was conducted.

A preliminary tryout was conducted with students from Ivy Tech on the Shelbyville and Bloomington campuses to gauge whether the students would be able to accomplish the task, to see if student reactions to the assessments would be positive, to identify any incorrect or confusing information in the text used, to gauge the clarity of the instructions, and to determine the amount of time students would need to read the material provided and to respond. These

preliminary field tests helped establish adequate time limits for the assessment and identified minor confusions caused by texts and graphics.

Ivy Tech was selected because of the emphasis in *Indiana House Bill 1695* and *Senate Enrolled Act No. 419* on vocational training and because the Ivy Tech administrators expressed keen interest in this type of assessment.

These preliminary trials revealed the need to change some language in the directions and to clarify some of the assessment procedures. After these adjustments were made, the major field trial was conducted.

The schools scheduled their own participation.

Teachers, administrators, and students from the schools which ended up participating in the field test were tremendously supportive and cooperated with patience and good will. Taking part in an additional large-scale assessment, however, was not a minor challenge for some of the schools scheduled to participate.

Teachers, administrators, and students...were tremendously supportive....

All schools participating were allowed to schedule their own testing dates, which spread consequently over a period of several months. Many schools scheduled the trial on regularly scheduled half days for testing; others scheduled the assessments during regular school time.

One school administrator protested that an adult would not be asked to read, assimilate, and respond to this amount of new information in a single three-hour session. He was among those who recommended that the assessment be split into two periods. Some teachers recommended that the assessment be scheduled over two or three days. Interestingly, few students complained about the timed arrangement for responding, although some teachers administering the

tests felt that some students just stopped working when the time requirement seemed burdensome to them.

An effort was made to make the directions as explicit as possible yet sensible. Survey results indicated that few teachers had any problems with the directions. Some teachers felt that while they approve of the assessment format and emphasis, they need some in-service training in order to adequately administer such assessments and to apply the results to their teaching.

The intention in designing the trial was to try out all ten assessments as evenly as possible across the different grade levels. Each of the assessments was administered to a substantial number of students; the numbers across grades who took each are somewhat uneven.

In asking about and examining the assessments, school officials and teachers began to express preferences for which assessments they would administer to whom. Some felt certain ones were more appropriate for their students, more appropriate for particular grade levels, or better instruments for revealing the kind of performance that interested them. Some schools indicated that the literature-based assessments would be best with their younger students and the workplace assessments would be more useful to their older students and serve their need to think about what employers look for.

...school officials and teachers began to express preferences for which assessments they would administer to whom.

The Center staff did not want to curtail such reactions. Therefore, the goals for balancing the number of subjects for all the assessments across all grade levels and types of students were relaxed and differences were tolerated as long as the totals were adequately large for each assessment.

Scoring System: Establishing consistent scoring for a wide range of individual responses

It is reasonable to ask: When students are responding in individualistic writing styles and using the information from several sources to wholly construct an answer, how can such responses be scored reliably? Won't the different scorers who assess these papers have preferences that vary just as much as the students' writing styles and opinions?

It should also be noted again that responses that go beyond the reading prompt and incorporate individual experiences are to be valued and encouraged on performance testing. That is, after all, what reading and reacting are all about—adding the meaning constructed by reading to what one already knows as a new synthesis.

Reliable methods of scoring have been developed to account for such differences. Such scoring methods can standardize the scoring of student writing and thinking without negating the kinds of values that must be appreciated subjectively. It is possible to construct sets of criteria that are clearly enough articulated that they will guide different raters who rely on them to assign the same or very similar scores to the papers they read.

Reliable methods of scoring...can standardize the scoring of student writing and thinking without negating the kinds of values that must be appreciated subjectively.

In addition to designing the assessment prompts, tasks, and writing aids, a key task was to develop viable rubrics (scoring criteria guidelines), to use these to identify anchor papers that could exemplify each score for each dimension, and to create a scoring manual using the rubrics and anchor papers.

In this presentation, the explanation of the criteria for each possible score were, therefore, criterion-referenced to the anchor papers.

Reliable scoring criteria had to be created.

The development of the scoring systems for both the reading-writing and math-communications assessments took place over several months and emerged from a complex multi-step process. The scoring system that resulted from the process needed to achieve several general goals:

- Encompass those attributes noted in research and seen by teachers as highlighting superior papers,
- Be clearly related to the identified skills,
- Transcend the individual genres, and
- Be easily understood by both students and those who would administer the assessment.

Specific rubrics, on the other hand, were also essential and needed to be:

- Specific to the assessment,
- Able to accommodate both the unique requirements of the task and general scoring guidelines, and
- Capable of guiding consistent evaluation across raters.

In the first of several steps, the developers used research and theory available on holistic scoring to outline the scoring *dimensions* and several key indicators or factors that could help define or describe those dimensions. This would serve as an initial draft of a generic rubric. The three dimensions for the reading-writing assessments (*Accomplishment of Task, Reading,* and *Writing)* had been selected early in the project to guide the construction of the assessments. The scoring range had been selected as well—a scale of *1* to *4*.

With the scoring range and dimensions in place, completing the scoring system was a matter of describing the scores for each of the four scores across

the three dimensions. The differences between a *4* and a *3*, between a *3* and a *2*, and between a *2* and a *1* had to be described clearly and in adequate detail.

First, key factors which helped identify each dimension were identified. While it was not intended that these be scored separately, they are guides to considering each dimension and are included in the rubric within the description of each possible score. If three particular factors are used to help define a score of a *4* on *Writing*, for example, the same three factors must also be considered in describing the other three scores below it. Something different about each of these factors will distinguish each of the four scores from the one above and the one below it. Whatever words (degree, frequency, quality, etc.) describe the factor, they must be greater or better than those that describe that factor for the score below and smaller or less than those used in describing the score above.

Factors used to help define each dimension included these:

- If reading ability is evident in the writing, it was decided, the response will demonstrate a clear understanding of the directions, will use appropriate material or details selected from the reading passages, and will give evidence of an understanding of the main concepts in the prompts.
- Writing ability was defined in terms of word choices, sentence structures, attention to language conventions, and the creation of clear relationships between portions of the written response.
- How well the task is accomplished can be judged, it was decided, by the organization of the writing, by its main points and the details that support them, by attention to audience needs, and by the inclusion of ideas and information contributed by the writer.

These outlines were graduated as a generic rubric to show four levels of quality, from the extremely well done task to the totally inappropriate response.

The definition of scores in the rubric were then compared to the instructions given to the students, and teachers were asked to offer suggestions and corrections. The entire Center team looked at, practiced applying, and worked at revising the rubrics at this stage.

The rubrics were then held as final until after the administration of the assessments to the entire pool of field trial subjects. At that point, each developer selected one assessment and attempted to score and rank 150 to 200 student responses. The rubrics were subsequently reanalyzed in terms of variance in scores assigned across raters to identify and eliminate alternative interpretations of assessment criteria. The generic rubrics were revised and polished one more time to become a part of the scoring guide for each assessment.

The procedure for this step in developing the scoring, involved the following activities: Members of the development team each read a sizable number of papers for each assessment. Individually, they sorted them into piles of those that they felt were quite strong, those they felt were obviously inadequate, and those that could not be put clearly into either pile. This was ultimately repeated for each of the dimensions in the generic rubric. Then the strong and weak papers were examined to identify and note how factors within the dimensions were commonly articulated throughout the pile.

Next the developers met and compared their results. This led to extensive discussion of the specification of factors and how they were distinguished across the sets of responses. From this effort, the team was able to draft a specific rubric for each assessment.

Model (anchor) papers were selected for use in scoring manuals.

The final step in rubric construction was to use them to develop the anchor papers. The entire development team worked together to identify

example papers that were needed for each assessment to reflect each dimension at each level of quality. Using the applicable rubrics, the Center staff each took a number of different student papers for each assessment and read them in order to identify examples of each of the four scores in each of the three dimensions. Piles of the papers coinciding with the four score points for each of the three dimensions that were scored were brought together and discussed until the team identified two examples for each score possible on each dimension.

These were *not necessarily* the best two in the group's opinion. Rather, the team tried to find papers with the same score on a particular dimension that were distinctly different in some way. Perhaps one was very strong in all factors a solid *4*, and the other was a *4*, but with far less clear credentials. Perhaps one was very strong in the first, second, and fourth factors but was only average on the third; and the other anchor paper selected was unusually strong on the second and third factors and was only above average on the first and fourth. Ultimately each pair of anchors for each possible score in each rubric were *randomly* labeled *a* and *b*.

A final and very important part of creating the anchor papers is the writing of annotations that explain how, in terms of the rubric, each paper was given the score it represents for a particular dimension. These descriptions are needed to cite specific examples in each paper to guide the future rater's use of the anchors in assigning scores to other student efforts. The set of annotated anchor papers for each prompts was reproduced in a Book of Model Papers. Copies of these books are included in the appendix.

annotations...explain how...each paper was given the score it represents for a particular dimension.

In contrast to the procedures used with the reading-writing prompts, each math-communications model paper was scored for four dimensions:

Reasoning, Knowledge of concepts, Communication, and *Use of procedures.* Therefore, each model included was reported with all four scores: for example, a paper would exemplify the scores *4, 3, 3, 4*; *1, 1, 2, 1*; or other combinations.

Six to ten models were used to develop a Mathematics Book of Model Papers for each assessment. Each paper was introduced with a short annotation which explained the reasons why each score was assigned to each dimension. The goal of the annotations and collection of anchor papers for each math-communications prompt was, as in the reading-writing assessment model papers, to guide a rater's scoring of other papers.

With the rubrics, the collections of anchor papers comprised a scoring manual for each of the ten reading-writing and math-communications assessments that was more like a reference manual than an answer key. Each manual began by presenting an overview of the theory underlying performance assessment, background material about the assessment development process, general guidelines for holistic scoring, and specific guidelines for analytic scoring. Then the rubrics and model papers were presented as reference points for the scorer to use.

Reliability: How the reliability of the scoring system was determined

Throughout this study there was a continual emphasis on the creation of assessments and rubrics that could be handled by local school districts without compromising any of the needs of a statewide assessment. For such assessment to meet the needs of audiences interested in accountability, for example, it must be *reliable* and an *authentically valid* example of what we teach and expect our students to be able to do. To be reliable, a scoring system has to guide different raters using the scoring guide to the same, or nearly the same, scores for the same student responses.

This does not require experienced raters; teachers who are inexperienced in evaluating performance assessment results should be able to do this. One key question of this study was *What kind of training is essential to achieve this result?*

Therefore conducting workshops to train teachers as scorers was considered an integral portion of the project. Workshops were conducted in January, March, and April of 1993. Educators who had participated in the field trials were given the first opportunity to attend. A total of 54 participants were trained: 40 teachers at the reading-writing assessment workshop and 14 at the math-communications workshop.

Those who wanted to be in the scoring workshop were required to read and work through the assessments for two of the ten assessments before they came to the Bloomington campus. This was to make sure that they came to the workshop with some familiarity for the assessment procedures and format, as well as for the material used in the two prompts.

The participants covered a wide range of teachers, many of whom had never done any kind of holistic scoring before. They were not prescreened beyond the fact that their schools had participated in the field trials and that they expressed interest in learning more about performance assessment. Some were more experienced at evaluating student writing and mathematics than others. Some had never taught English or assigned and graded much student writing.

Only a half day of each two-day workshop was available to train the scorers, so that they would have a day and a half to apply and practice what they learned. The first portion of the workshop was devoted to acquainting the scorers with the philosophy behind holistic scoring conducted with criteria indicators. Next the participants were given several training papers that had been previously scored. They were asked to score these papers according to the rubrics and using the booklets compiled for each assessment with annotated anchor papers for each possible score under each of the dimensions.

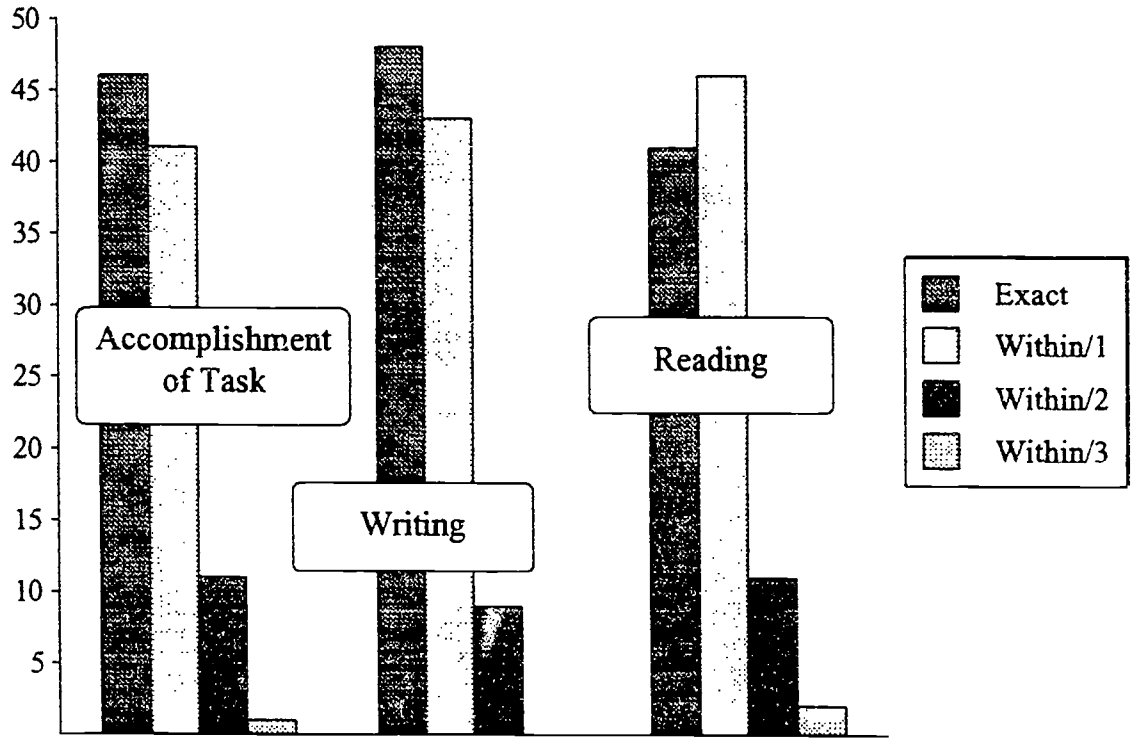
There were 24 annotated model papers for each reading-writing assessment and a range of seven to ten for each of the math-communications assessments to assist the scoring of responses to each prompt. During this initial session, the scorers began using these and familiarizing themselves with them. As the members of teams that were formed among the workshop participants did this early scoring, they discussed why they had given each score for each dimension. Their scores were compared to those of the previous scoring, and the discussion turned to a focus on differences.

...participants were able to come within one score point of a second scorer on 80 to 90 percent of the scores assigned.

As a result of this workshop, participants were able to come within one score point of a second scorer on 80 to 90 percent of the scores assigned. While the ideal scoring result would be for all scorers to agree on the same score, it

was not the goal of this workshop to develop this high a reliability. The time available was not sufficient to strive for this degree of reliability. Coming within one score point was a more realistic goal, and what was learned about training from this part of the study could ultimately contribute to training that would produce higher reliability.

Chart 1
Percent of Scoring Agreement After Minimal Training
California Water Prompt

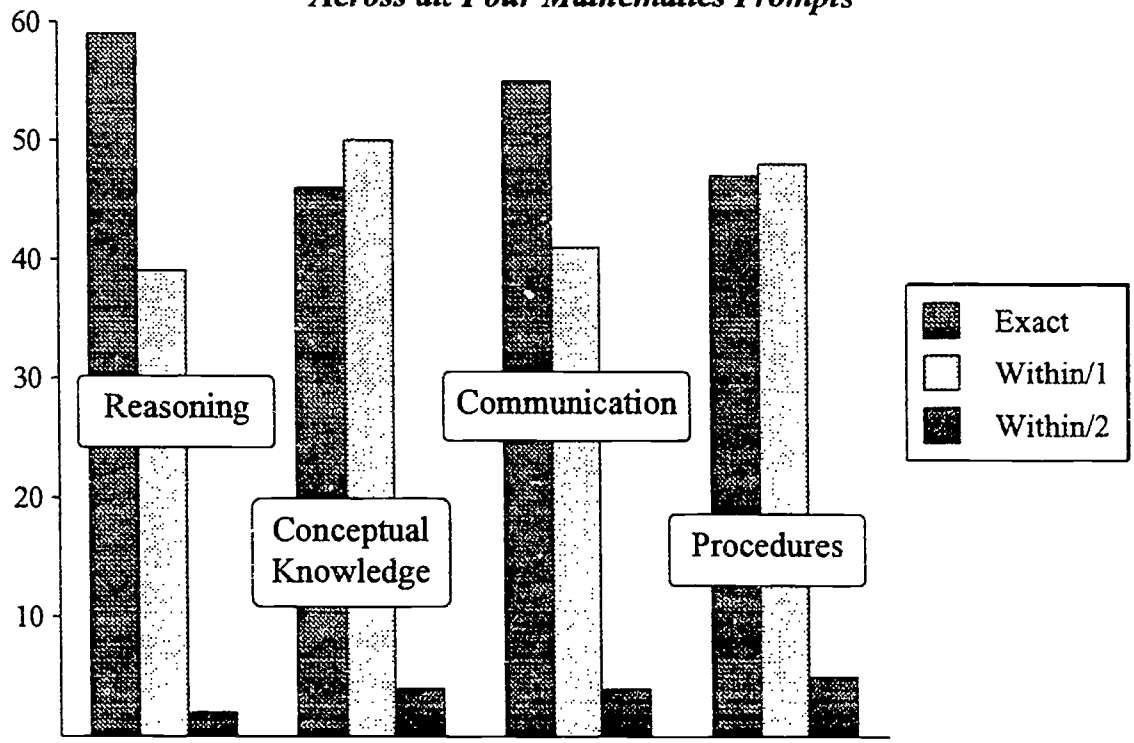


The number of papers that received exact scores, scores plus or minus 1, and scores plus or minus 2 are shown in Chart 1, which shows the scoring agreement achieved with this minimal training for one of the assessments.

Math reliability: The reliability produced by matching scores assigned by two scorers to 107 papers in the mathematics workshop were as follows:

- Exactly the same scores: From 46 percent to 59 percent across the four dimensions;
- Plus or minus one point: From 39 percent to 50 percent;
- Plus or minus 2 points: From 2 percent to 5 percent.

Chart 2
Percent of Scoring Agreement After Minimal Training
Across all Four Mathematics Prompts



Thus, there was agreement within one point on from 95 percent to 98 percent of the scores assigned, depending on the dimensions for correlations that ranged from .64 to .75. The highest correlation was achieved in scoring *Reasoning.* Chart 2 depicts these results.

Participant feedback: A significant goal of these workshops was to receive feedback from teachers exposed to this form of scoring in order to obtain more information about the viability of the rubrics. Scorers completed a short questionnaire about the scoring workshop, about the reference and

resource materials they were taught to use, and about the overall experience. They agreed that they could score the responses easily (89 percent) and quickly (86 percent). They indicated that the rubrics (79 percent) and model anchor papers (97 percent) helped them in scoring. Almost two-thirds agreed that teachers would use this type of assessment to change instruction.

The participants did, indeed, in their discussion of variance across their scores, identify points in the description of the criteria that have been targeted for reconsideration and possible revision. Equally important, participants shared with the Center team ideas they had about what specific training scorers will need, potential new topics for assessments, and other aspects of the project.

Still another objective of the workshop was to identify a cadre of teachers interested in willing to learn more about performance assessment, and particularly about scoring it. The workshop provided the opportunity to disseminate information and training to these persons from various parts of Indiana and to enlist their commitment to any state initiative of this type.

Can intensified training produce higher inter-scorer reliability?

The workshop for training scorers of the reading-writing assessments was followed by a study which attempted to achieve a higher correlation (reliability) between scorers. The findings from this study underline the importance of in-depth training of scorers and indicate that higher reliability is possible with intensified training.

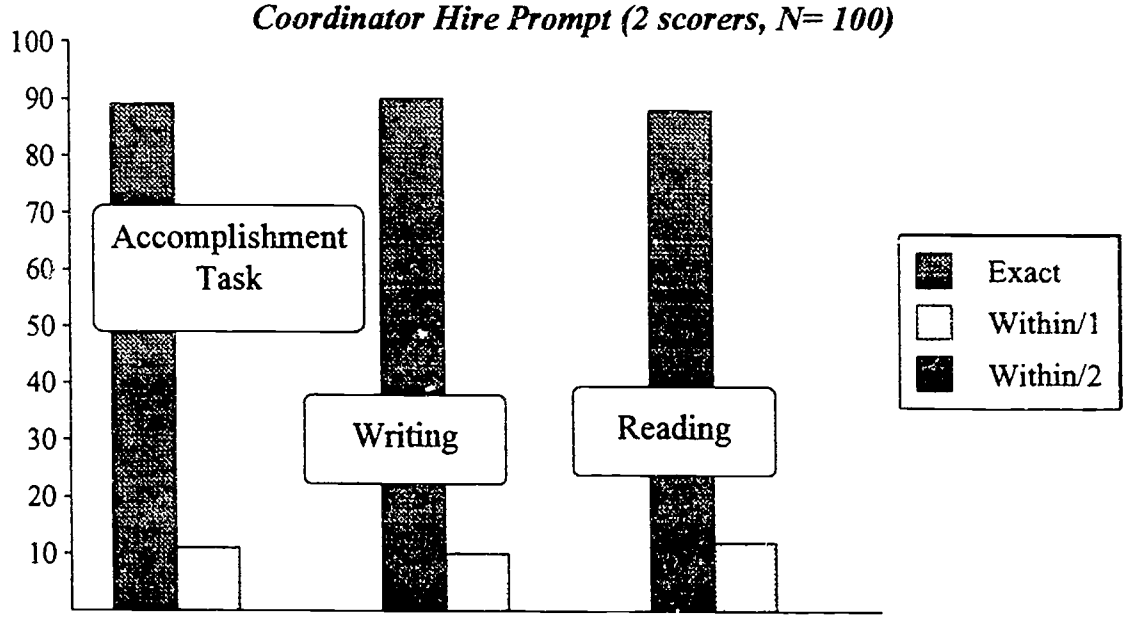
Four scorers agreed to participate in more extensive training designed to rely more directly on their collective discussions. This workshop focused on the reading-writing assessments. All the papers used in this workshop were considered without reference to any score or scoring notes from the previous workshop. The scorers began by working together to jointly score and discuss several papers. They analyzed the scoring criteria on the rubrics and the

annotated anchor papers selected to exemplify the scores described on the rubrics. This group analysis considered the relative importance of each factor as a reference point in determining the score for the dimension under which it appeared.

After this intensive self-training step of this follow-up workshop, this group scored 125 papers from the integrated content area and 100 from each of the other two genres of reading-writing assessments.

The reliability for the *Coordinator Hire Prompt* was the highest, and that for the *Relationships Prompt,* the assessment which asked students to write an introduction for the creative writing magazine, was lowest. But the reliability *within* one point was high across all three genres and perfect on the *Coordinator*

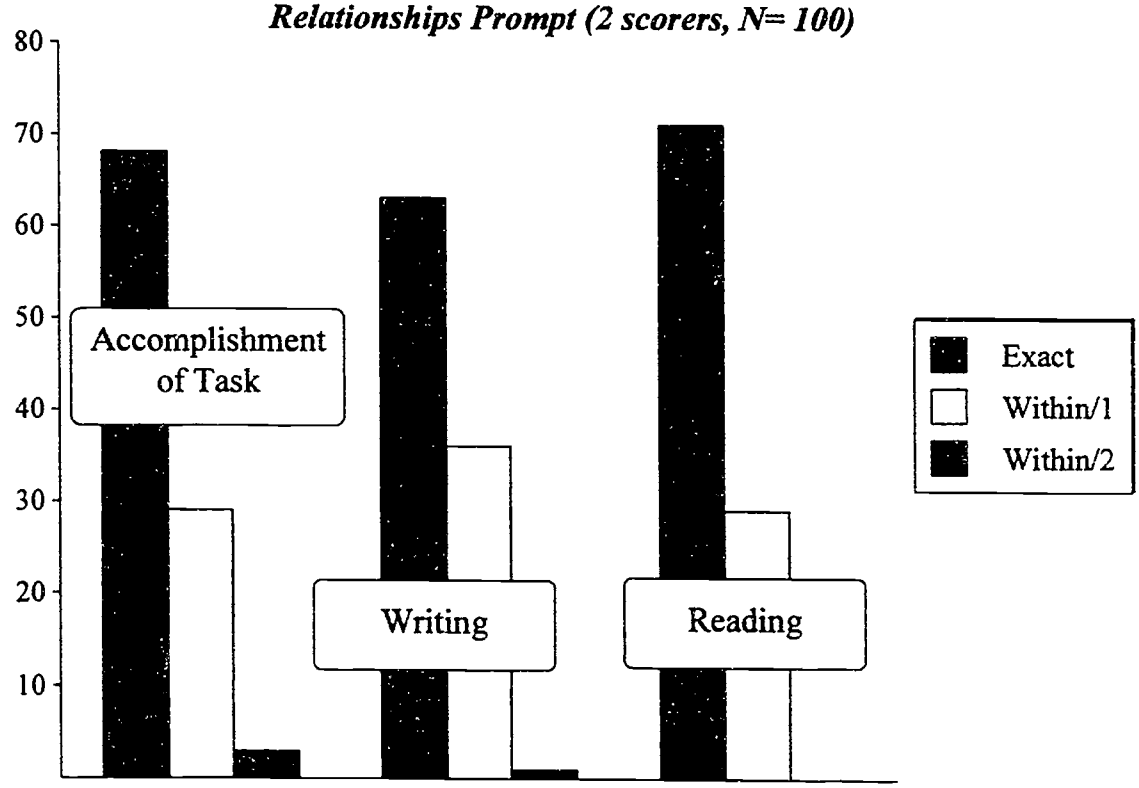
Chart 3
Percent of Scoring Agreement After Intensive Training
Coordinator Hire Prompt (2 scorers, N= 100)



Hire prompt. Charts 3 and 4 indicate the range of reliability achieved with this more intensive training. When compared to the scoring agreement from the first reading-writing scoring workshop, these data show an increase in agreement

within one point of 97 percent to 100 percent. Equally important, they show a shift from agreement within one point toward exact agreement.

Chart 4
Percent of Scoring Agreement After Intensive Training
Relationships Prompt (2 scorers, N= 100)



This workshop demonstrated the importance of more intensive training in increasing inter-scorer reliability. It also suggests that in about one percent of the scores that are assigned, there is scorer disagreement of over one point. In large-scale, high-stakes assessment using such a system, papers with such scorer disagreement could be rescored for fairness to the student without creating a heavy burden.

Several conclusions were drawn from conducting this workshop.

This workshop demonstrated that teachers are interested in, and capable of, applying holistic scoring with analytic reference points. The methodology

discussions in the workshops and from the trial itself that a great deal of practical data is derived from this kind of assessment. In the opinion of the participants, all the assessments seemed viable, and the assessments created for the integrated content areas gave the best indications about student ability and promised to be the most useful instructionally. This reaction of the participants of the intensive workshop appears to endorse what was found in the earliest tryouts with post-secondary students, which indicated that both teachers and students felt that the assessments yielded useful data for instruction and personal growth.

In the opinion of the participants, all of the assessments seemed viable....

It was also concluded from the workshops that students and teachers would get a fairer report on student performance if the scores assigned by two scorers were combined and reported on an eight-point scale. Therefore, 4,000 of the papers from the trials were scored at least twice, using the four-point scale on all dimensions—three for the reading-writing prompts and four for the mathematics-communications prompts. The two scores given for each reading-writing dimension were combined and reported on the eight-point scale. If one scorer gave a paper a *3* for *Reading,* for example, and another gave it a *2*, the score would be reported as a *5* out of a possible *8*.

The scores were reported to the participating schools using the eight-point scale. The interpretation of these scores to the schools was accomplished in two ways. First, to account for the doubling of the scores, the anchor papers in the scoring booklets were labeled as follows:

- An anchor paper of *4* was labeled as an *8*.
- An anchor paper of *3* was labeled as a *6/7*.
- An anchor paper of *2* was labeled as a *4/5*.
- An anchor paper of *1* was labeled as a *2/3*.

The school personnel who received the score reports were then told that they could turn to the books of model papers to determine exactly what each score represented in terms of a particular student's performance. If a student received a score of 6 on the dimension *Writing,* for example, the school personnel could turn the to the anchor paper labeled 6/7 and see exactly what that score represented in terms of student performance.

In addition, the project staff added their own qualitative interpretation of the scores. These qualitative interpretations were subjective and were based on the teaching experience of the staff as well as the experience of the staff in previous test development work. These qualitative interpretations were also reported to the schools. The interpretations were as follows:

- A score of *8* was considered to be *superior,*
- Scores of *6* or *7* were considered to be *very good,*
- Scores of *4* or *5* were considered to be *satisfactory, but they indicated the need for additional instruction,* and
- Scores of *2* or *3* were considered to be *poor* performance.

Very few papers were deemed "unscorable" because they were such inappropriate responses, so it was necessary to define the *0* to *1* range on the eight-point scale with this term. Accompanying each report was a booklet with the scoring criteria rubrics and the anchor papers for each assessment given at the particular school as criteria reference comparisons.

Findings and Discussion: How did students perform on these types of assessments?

A great deal was learned from this project about the prospects for using performance instruments and methodologies of this type in large-scale assessment. Most importantly this project determined how to develop, administer, score, and interpret these types of performance assessments. The fact that this approach to holistic assessment can be made reliable has been demonstrated and reported above. In addition, the viability of such assessment would, of course, depend on what it reveals about student performance. Those findings are reported in this first section on results. Teacher and student reactions to the assessments are reported in the following section.

What did the reading-writing assessments reveal about the performance of Indiana students?

The field trial clearly demonstrated that Indiana students are capable of reading, comprehending, and responding well to this form of assessment. With the exception of a few papers on which students responded so obliquely to the task that raters deemed them unscorable, virtually every student was able to understand the task and to respond to the prompt.

The data in Table 1, *Mean Scores by Grade Across All Prompts*, shows that the higher the grade, the higher the mean score—as would be expected if the test is validly assessing behaviors that are taught and should develop from one grade to the next. The difference between Grade 11 scores and those for the I.U. seniors is large, perhaps indicating the fact that a college senior group is a much more selective group than high school students. *Writing* scores tended to be closer to *Accomplishment of Task*

scores than were *Reading* scores, which were lowest among the dimensions for the high school students. For the college seniors, however, *Reading* was higher than *Writing*. The data reported in Table 2, *Mean Scores for All Grades by Prompts*, describes the mean score differences across all of the reading/writing prompts. The results indicate that there were not large differences between the mean scores for each of the prompts. However, the *Purchase* prompt seems to be the easiest and *Relationships* prompt to be the most difficult of the assessments.

Table 3, *Mean Scores by Dimension within Grade for Each Prompt,* gives the mean scores for each of the three dimensions within each assessment at grades 9, 10, and 11 and for a group of seniors at Indiana University, who also completed the *Water Problem* assessment.

It is interesting, as *Table 3* demonstrates, that across the grades and different assessments, it is not possible to conclude that the students always performed better within one dimension than within the others. Their performance differed by dimension across prompts. The issue of comparability of prompts is an important issue in the further development of these types of assessments and is continuing to be studied.

The *Water Problem* and *Sports* assessments appear to have been more difficult for the students than the others. The students were most successful in responding with the memo about which bicycles to stock or which candidate to hire, both created as "real-world literacy" tasks. The performance of the grade nine subjects on the workplace-type prompt *Coordinator Hire*, however, was low in comparison to the other prompts. The grade 11 responses were strongest on the assessment about *Fuel Tax*.

Table 1: ***Mean Scores by Grade Across All Prompts***
Indiana Performance Assessments '92 **(Reading-Writing)**
(Reported as a combination of the rating of two scorers on an 8-point scale)

Dimensions	Grade 9	Grade 10	Grade 11	IU Seniors
Accomplishment of Task	4.29	4.46	4.99	7.16
Writing	4.25	4.45	4.90	6.91
Reading	4.16	4.38	4.86	6.97

Table 2: ***Mean Scores for all Grade by Prompts***
Indiana Performance Assessments '92 **(Reading-Writing)**
(Reported as a combination of the rating of two scorers on an 8-point scale)

Dimensions	Rela-tionships	Water Problem	Fuel Tax	Hire	Purchase	Sports
Accomplishment of Task	4.58	4.24	4.52	4.75	5.17	4.25
Writing	4.74	4.26	4.45	4.64	4.91	5.34
Reading	4.53	4.03	4.59	4.80	5.09	3.85

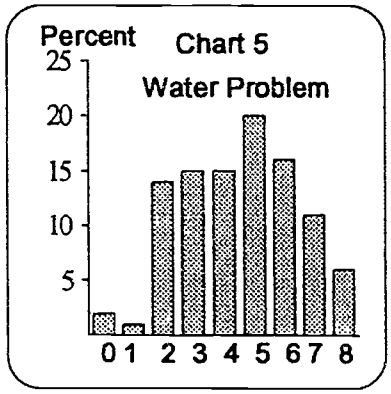
Table 3: ***Mean Scores by Dimension Within Grade for Each Prompt***
Indiana Performance Assessments '92 **(Reading-Writing)**
(Reported as a combination of the rating of two scorers on an 8-point scale)

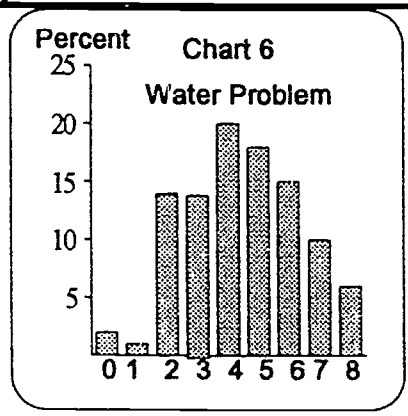
Dimensions	**Rela-tionships**	**Water Problem**	**Fuel Tax**	**Hire**	**Purchase**	**Sports**
Grade 9 Scores	*N=180*	*N=147*	*N=401*	*N=31*	*N=112*	*N=157*
Accomplishment of Task	4.31	3.86	4.17	2.77	5.15	4.64
Writing	4.34	3.90	4.11	3.06	4.88	4.59
Reading	4.22	3.57	4.27	2.87	5.21	3.86
Grade 10 Scores	*N=182*	*N=388*	*N=83*	*N=111*	*N=179*	*N=132*
Accomplishment of Task	4.46	4.03	4.81	4.69	5.21	4.29
Writing	4.60	4.09	4.75	4.61	4.91	4.33
Reading	4.42	3.87	4.80	4.83	5.23	4.04
Grade 11 Scores	*N=94*	*N=105*	*N=80*	*N=359*	*N=206*	*N=126*
Accomplishment of Task	4.93	4.63	5.78	5.07	5.23	4.20
Writing	5.40	4.57	5.64	4.84	4.95	4.43
Reading	4.91	4.39	5.78	5.07	4.97	3.88
IU Seniors' Scores		*N=32*				
Accomplishment of Task		7.16				
Writing		6.91				
Reading		6.97				

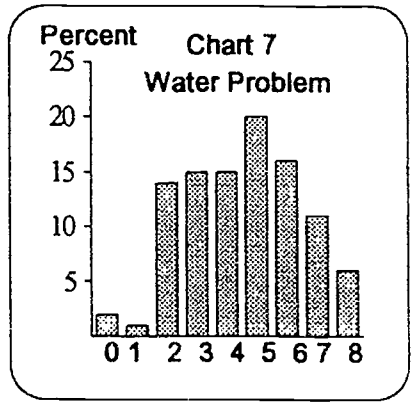
Charts 5 to 22 provide the percentage of particular scores on the eight-point reporting scale for the three dimensions of the reading-writing assessments by the different prompts. They demonstrate that the impact of the assessment content and/or perhaps the task required may have affected student performance.

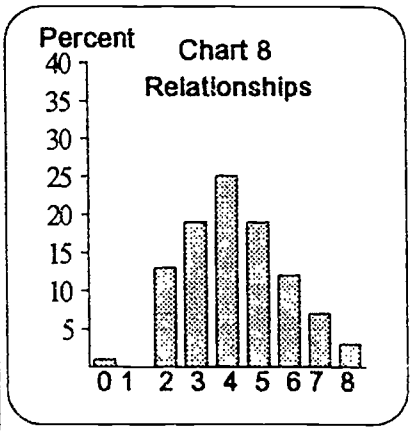
Charts 5-22: Students' Performance Scores in Percentages by Dimensions for Each of the Reading-Writing Prompts
***Indiana Performance Assessments '92* (Reading-Writing)**
(Ratings of two scorers combined on an 8-point scale)

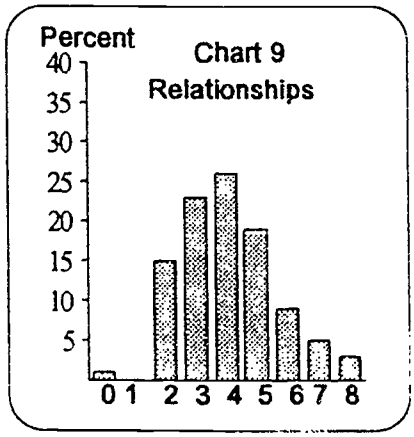
Accomplishment of Task	Writing	Reading

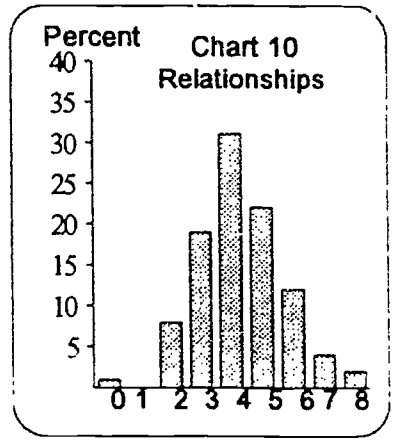




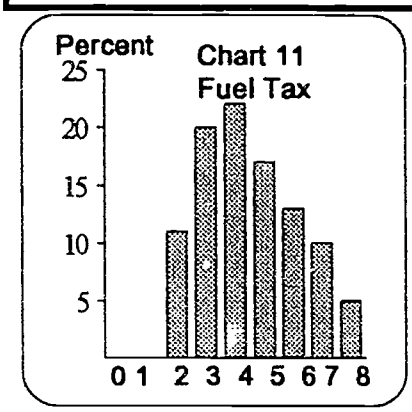


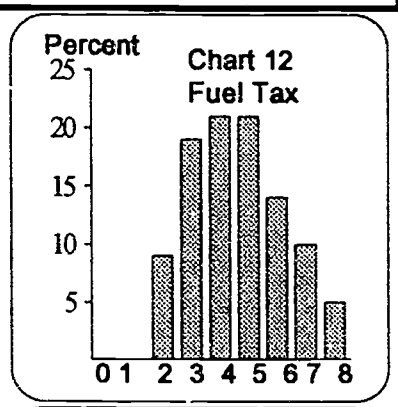


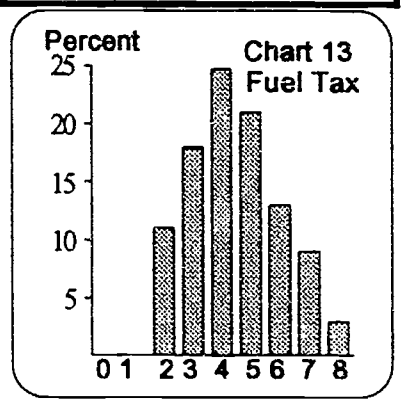


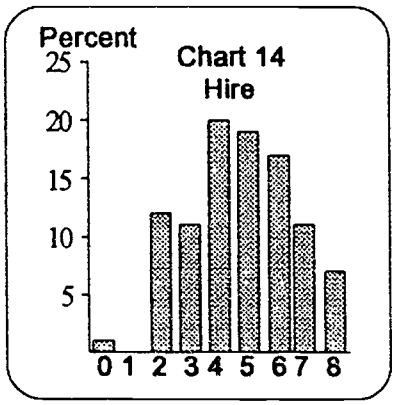


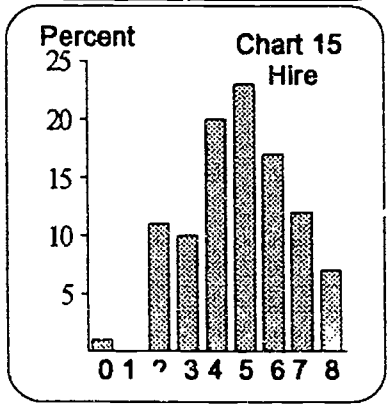
Accomplishment of Task	Writing	Reading

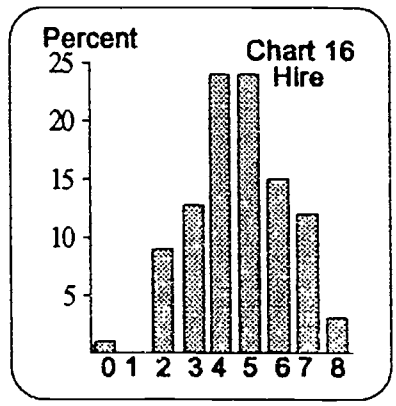


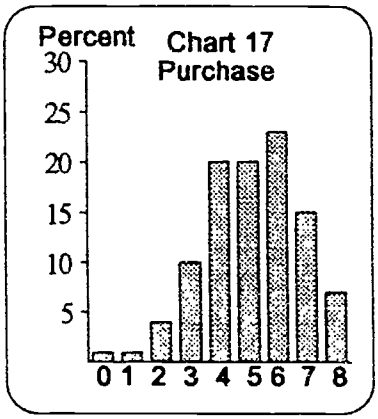


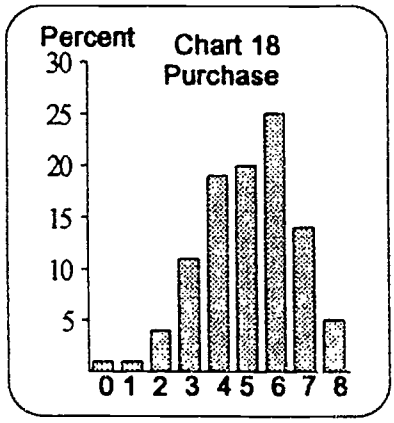


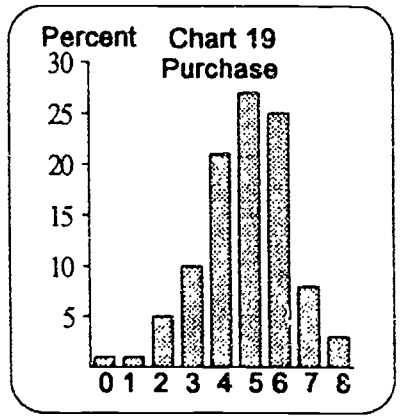


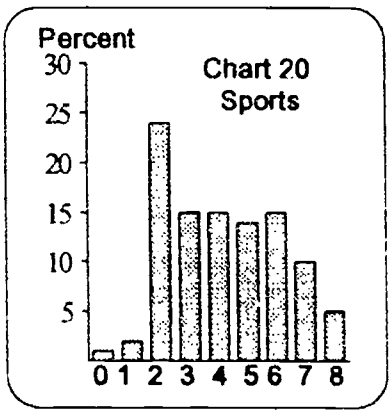


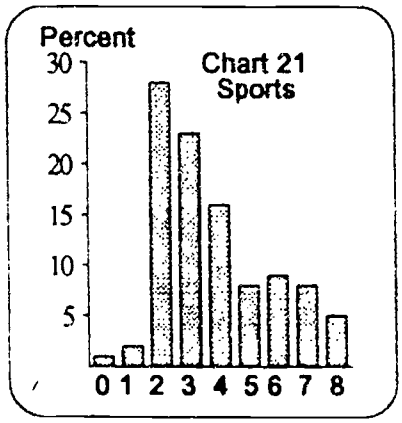












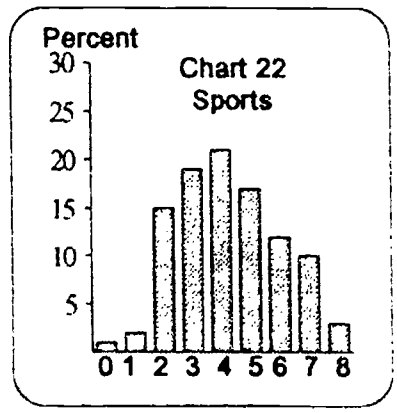


Chart 23 shows the reading-writing scores across all prompts for each of the grades. Chart 24 shows the expected difference in scores on the *Accomplishment of task dimension* for Grades 9 and 11 on the reading-writing assessments across all prompts.

Chart 23: ***Scores Across All Prompts for Each Grade***
***Indiana Performance Assessments '92* (Reading-Writing)**
(N=3,373)

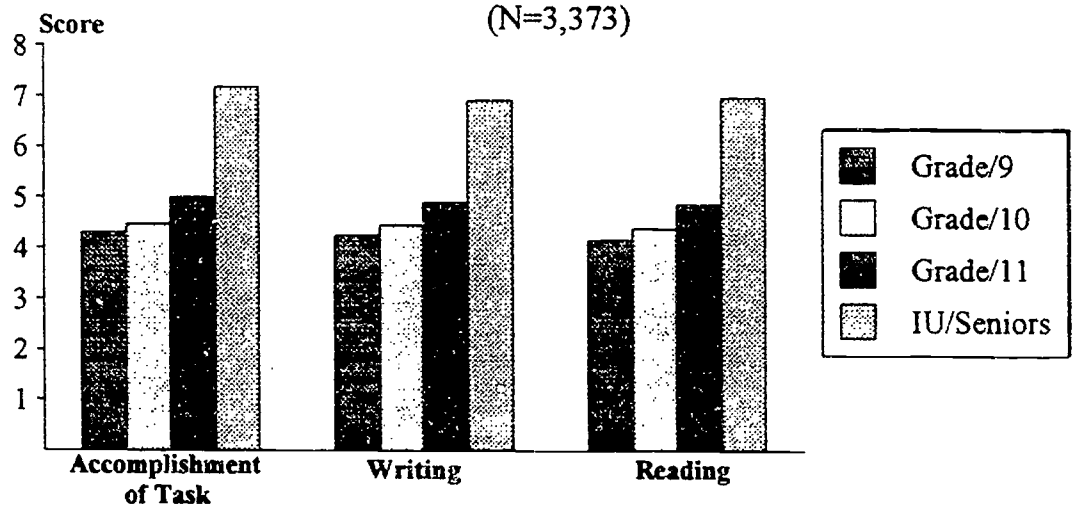
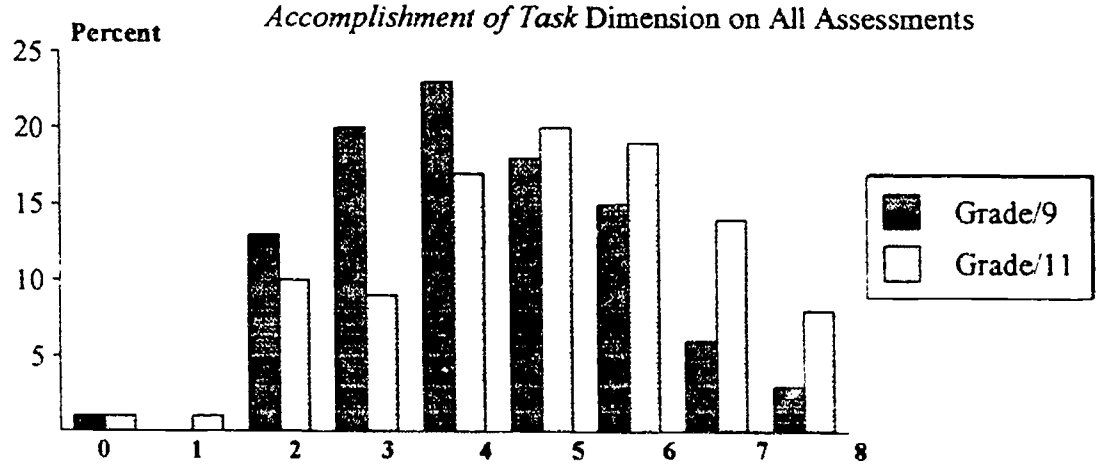


Chart 24: ***Comparison of Scores***
Grade 9 (N=1,028) and Grade 11 (N=970)
Accomplishment of Task Dimension on All Assessments



Pilot study data must be interpreted cautiously. In general terms, the scores for particular dimensions on each prompt were quite similar: Almost every mean value falls between 4 and 5, the midpoint of the eight-point scale. Careful examination of Charts 5-22 reveals several patterns of results. For example *Hire* looks very much like the normal curve; *Water Problem* appears rather flat across the range; *Purchase* is skewed to higher scores; and *Relationships* is skewed to the lower scores. However, one needs to consider the number of students who took each prompt, the distribution of grade levels, and the pilot nature of the administration when interpreting the results.

What did the mathematics-communications assessments reveal about the performance of Indiana students?

Performance on the mathematics-communications tests were skewed more toward the lower scores than were the reading-writing assessment results. At grades 9 and 10, more than half of the students scored a *1*. About 35 percent of the grade 11 students scored *1*. Table 4 reports the number of papers assigned each score *on the 4-point scale*. It should be noted that a significant number of papers were determined to be unscorable because students had failed to follow the directions.

The math performance assessments indicate that these assessments were more dissimilar to the kinds of activities that students engage in during mathematics instruction. The students commented that the assessments were difficult and that they were unfamiliar with the tasks. Despite the low scores, both students and teachers commented on the surveys that the mathematics assessments were focused on important outcomes.

Table 4: ***Number Receiving Each Score on All Prompts Within Grades***
Indiana Performance Assessments '92 **(Mathematics-Communications)**

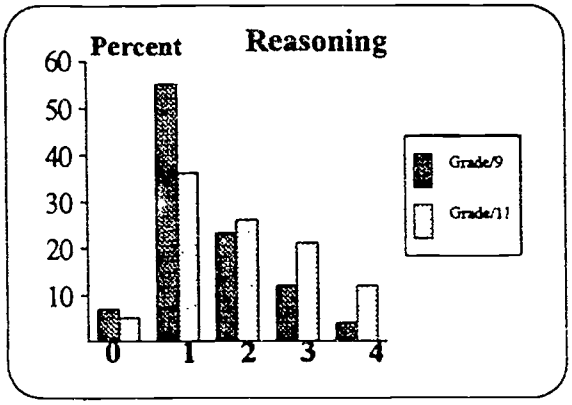
Score	Reasoning	Knowledge	Communication	Procedures
Grade 9				
0	50	69	31	66
1	412	396	293	380
2	170	175	296	181
3	91	81	99	96
4	29	31	33	29
Total	752	752	752	752
Grade 10				
0	65	63	57	58
1	370	341	306	328
2	185	218	233	219
3	125	139	135	147
4	71	55	85	64
Total	816	816	816	816
Grade 11				
0	48	52	47	48
1	351	341	326	338
2	251	263	244	258
3	202	200	195	199
4	116	112	156	125
Total	968	968	968	968

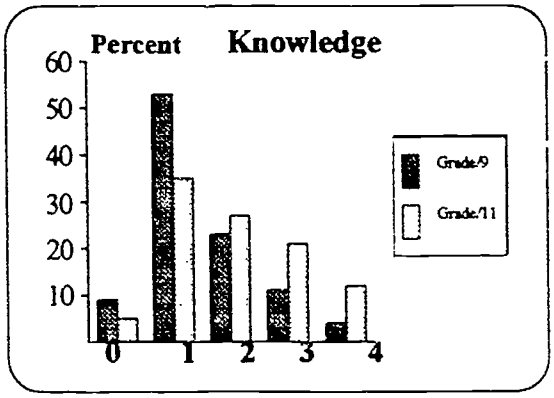
Charts 25-28 compare the scores for Grades 9 and 11 for all four assessments on each of the four dimensions, and clearly demonstrate the

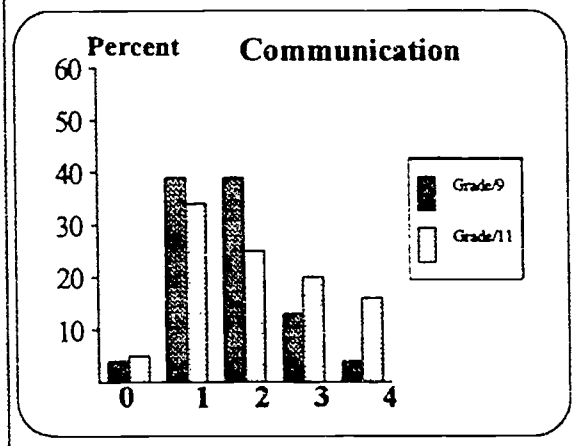
difficulty of the mathematics assessments. The data shows that the assessments are somewhat easier for the Grade 11 students, suggesting that what is being assessed may be an ability developed by additional instruction and perhaps maturation. Teacher and student reactions, reported in the next section indicate that both groups found the mathematics assessments to be difficult, and their reactions suggest some reasons for these difficulties.

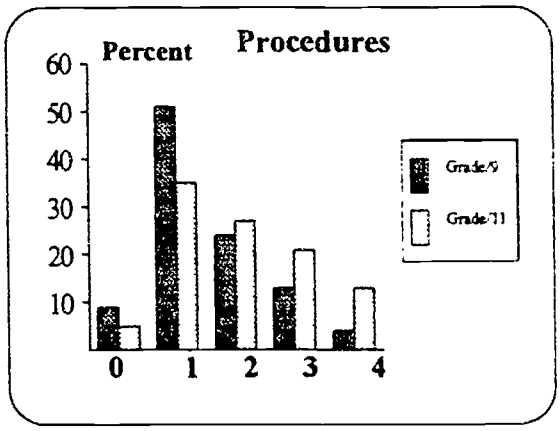
It has not been possible to determine the reason for the perceived difficulty. Two possible explanations are plausible: The assessment was hard, leading to rather low scores and to survey responses indicating this difficult; or the assessment was unusual and unfamiliar leading to confusion and to some amount of rejection by students who then did not work through the assessments. It is also possible that the students were simply unable to perform the tasks that had been developed. Further studies will provide a better understanding for the lower-than-expected scores and the survey responses stating that the assessment was difficult.

Charts 25-28: ***Comparison of Scores for Grade 9 and Grade 11***
For Each of the Four Dimensions on All Assessments
Indiana Performance Assessments '92 **(Mathematics-Communications)**
(N=1,721)









The results from this study match those released by the National Center for Education Statistics of the U.S. Department of Education and reported by the Associated Press the first of September, 1993, and using data created by the administration of a test in 1992 for the National Assessment of Educational Progress to nearly 250,000 elementary and high school students attending 10,000 schools in every state. That study used a test similar to the assessments used in the *Indiana Performance Assessments* '92; the NAEP test instructed grade 12 students: "This question requires you to show your work and explain your reasoning."

It was reported[6] that:

Only...nine percent of high school seniors tested could answer mathematics questions requiring problem solving skills.

The results show that students are "getting few opportunities to participate in problem solving in classrooms," said John Dossey, a visiting math professor at the U.S. Military Academy and former president of the National Council of Teachers of Mathematics....

"Here, we see what a student is capable of doing," Dossey said, adding that a multiple-choice test question with five potential answers gives the student a 20 percent chance of guessing the right response....Tests requiring such detailed answers [as those on the NAEP assessment did] are considered more effective in revealing what a student has learned.

[6] These comments are taken from Associated Press reports that appeared in *The Indianapolis Star* and other newspapers on September 1-2, 1993.

Reactions: Teachers' and students' survey responses to the assessments

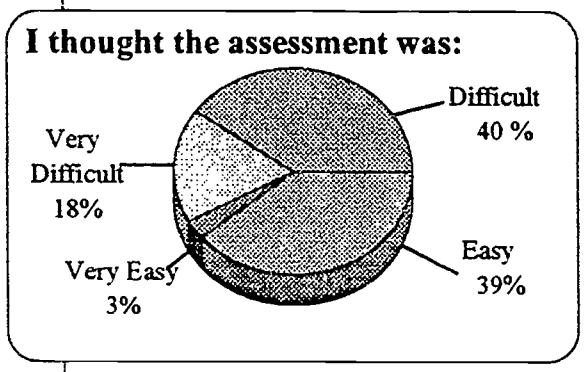
Any successful use of this type of assessment will depend in large part on its acceptance by students and teachers. Their reactions were recorded as the Center team worked with teachers from the trial sites, and they were solicited and recorded in the surveys conducted in conjunction with the trials. Those teacher and student reactions were clearly positive.

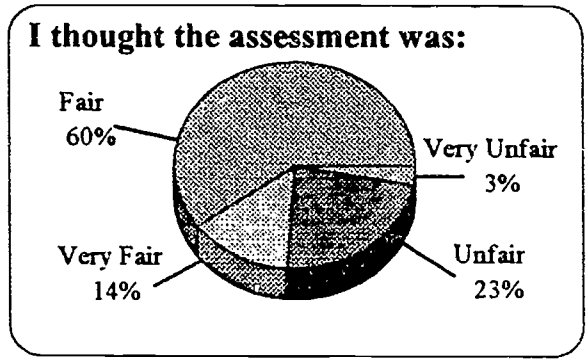
Over 80 percent of the students taking part in the field trials completed an 18-item questionnaire. Fifteen items had a four-alternative Likert-type scale, ranging, for example, from "very difficult" to "very easy" or from "very helpful" to "not helpful at all." Three of the questions had two or three alternatives. In addition, open-ended responses were solicited for questions that sought such reactions as those to the possibility of replacing traditional multiple-choice tests with the assessments being tried out or for "additional comments."

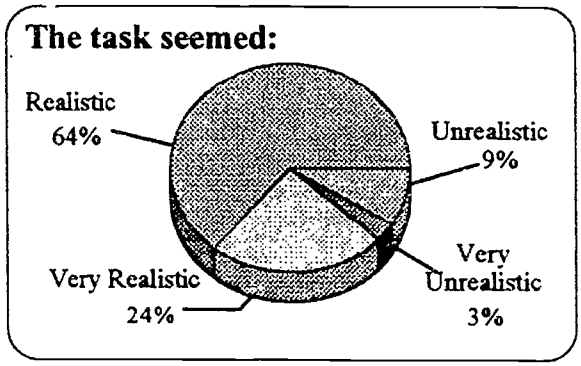
What did the survey reveal about students' attitudes regarding the reading-writing assessments?

The survey results tend to demonstrate balanced reactions that suggest that the assessments were of approximately the right difficulty: 58 percent of the students responding thought the assessments were "difficult" or "very difficult," and 42 percent thought they were either "easy" or "very easy." A high percentage, however, found that the assessments were "fair" or "very fair." The fact that 84 percent found the assessments "realistic" or "very realistic" tends to endorse the efforts to select topics and tasks that are *authentic* (realistic). Charts 29-31 depict these results.

Charts 29-31: ***Student Responses on the Survey:***
How Difficult, Fair, Authentic Were the Assessments?
Indiana Performance Assessments '92 **(Reading-Writing)**





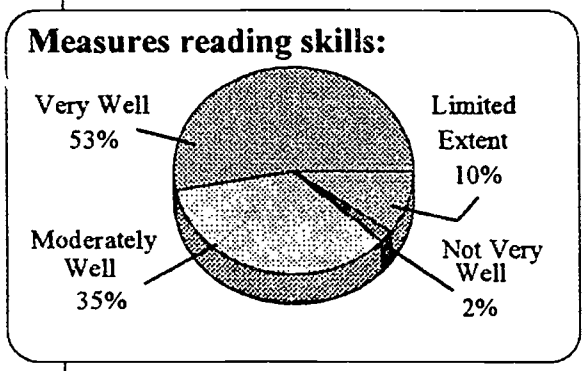


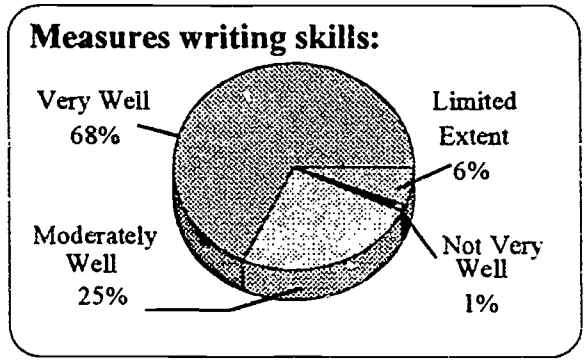
For assessments with such distinctively different content, the student reactions were very consistent across the six reading-writing assessments. At the same time, the students reported that the reading-writing activities assessed their reading, writing, and thinking abilities *very well*. Charts 32-34 show that when *moderately well* and *very well* are combined, about 90 percent of the students reported the positive reactions to the assessments.

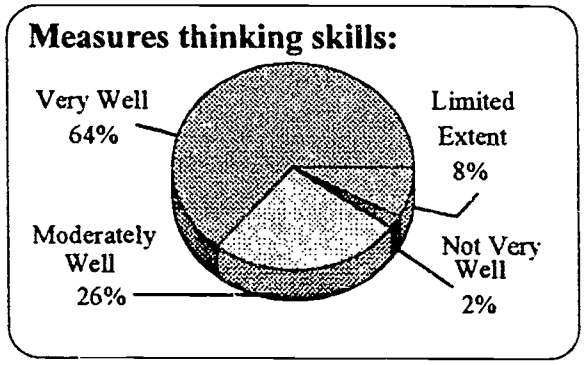
Other student reactions were equally positive:

- 65 percent reported that the assessment was very similar or somewhat similar to their experience and learning in school,
- 70 percent reported that their instruction had adequately prepared them for the test.

Charts 32-34: ***Student Responses on the Survey:***
How Well Was Reading, Writing, Thinking Measured?
Indiana Performance Assessments '92 **(Reading-Writing)**







Open-ended responses to the reading-writing assessments included these positive reactions to its replacing multiple-choice tests:

- It gives you something to think about instead of just guessing.
- It gets rid of guessing and requires some work.
- It's more interesting and showed our skills.
- People can learn as much writing as they do another way.
- This makes you think about your answer more.
- Your eyes get tired seeing the A, B, and C's [on multiple-choice tests].
- You get to use your skills in more ways.
- It makes it more realistic.

Negative responses to this question included:

- I hate writing papers.
- You get an idea what the right answer is with multiple-choice.

- One can't judge one's learning on only one work.
- I don't think people are ready to write an essay like this.
- You could figure out a grade easier with multiple-choice.
- Kids are not used to taking a test like this.

Other open-ended comments and suggestions included ones like these:

- You need to give topics to choose from.
- You should definitely give more time.
- At first I was confused but after reading and using the prewriting suggestion it was not at all that hard.
- It would have been more interesting if you had had a cop asking suspects about a murder, and....
- Give a choice of what kind of writing to use....
- I would allow four hours.

What did the teachers think of the reading-writing assessments?

Many teachers expressed pleasure in seeing an assessment that made the students think and, reason and many were curious about how the scoring would be handled. On the one hand, a few teachers expressed concerns about the length of the presentation booklet. But on the other hand, many praised the prewriting suggestions and attempted to get students to revise.

Numerous teachers noted the integrated nature of the test and how each of the six assessments required the students to engage in problem-solving. There were isolated criticisms of content, such as a challenge to the California water problem as not being appropriate for Hoosiers. Teacher reactions that were expressed by several or more individuals included these:

- With the exception of the *Relationships* prompt, which asked students to tie material for a student literary magazine together with an introduction, teachers found the assessments appropriately challenging—not too easy or difficult and apt to be meaningful for their students. From 67 to 94 percent

approved the other assessments, depending on the particular prompt.

- From two-thirds to three-fourths of the respondents found the directions for the particular assessments "very helpful" or "somewhat helpful."
- With the exception of the *Relationships* introductory essay, teachers thought their students had the kind of background to help them understand what they needed to do.
- While a few individuals recommended giving the students more time to think about the task at hand, from one-third to one-half thought the time allotment was too long. Most who commented on the time allotted thought that it was adequate.
- Teachers were positive about the experience for their students, about how well it matched what the students do in school, and about its potential to replace multiple-choice tests. Nearly all of them felt the assessments matched good instruction at least to some extent and, with the exception of the two literature-based genre assessments, measured reading ability either "moderately" or "very well." Over 90 percent thought all but the *Relationships* introduction task measured writing ability, and 67 percent said it measured thinking ability "very well."

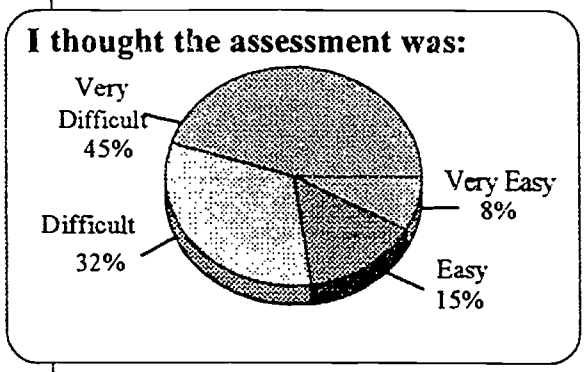
A table which details all the responses to the 12 questions on the survey by percentages for each prompt is in the appendix, as is a list of the open-ended comments from teachers responding to the survey.

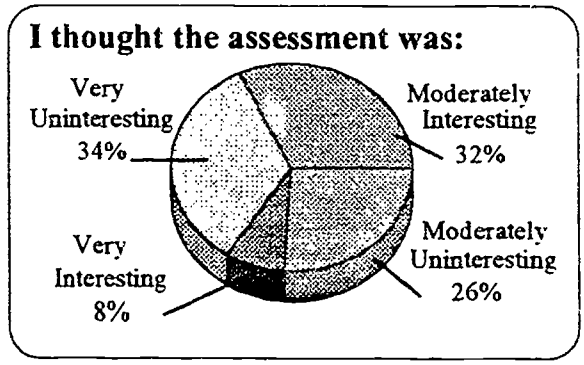
How did the students react to the mathematics-communications assessments?

While 66 percent of the students found the math assessments either "very interesting" or "moderately interesting," 77 percent thought they were "difficult" or "very difficult." Many students were aware, apparently, that they had not done well on the assessments, yet from 69 to 73 percent

reported that they thought the assessments had measured their problem-solving skills, their applications skills, and their computation skills "very well" or "moderately well." In these responses "very well" ranged from 30 to 46 percent. These reactions are depicted in Charts 35-39.

Charts 35-36: Student Responses on the Survey:
How Difficult and How Interesting Were They?
***Indiana Performance Assessments '92* (Mathematics-Communication)**





About half of the open-ended responses were positive, and included:

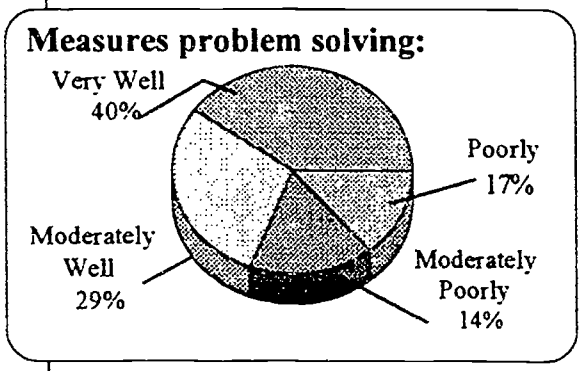
- It gives you an idea of what you're doing.
- It makes it more realistic.
- They are more interesting to do.
- It makes you think about it.
- You have to read the questions to understand them.
- Some questions have more than one answer.

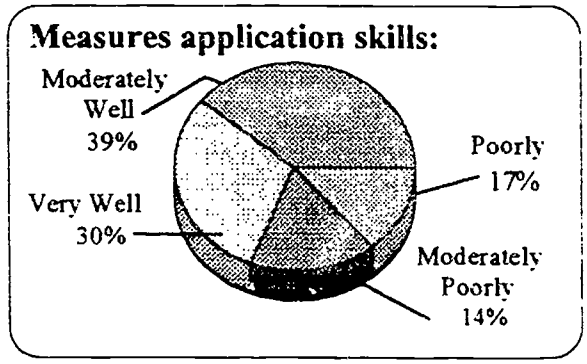
Some students were less positive about the assessment:

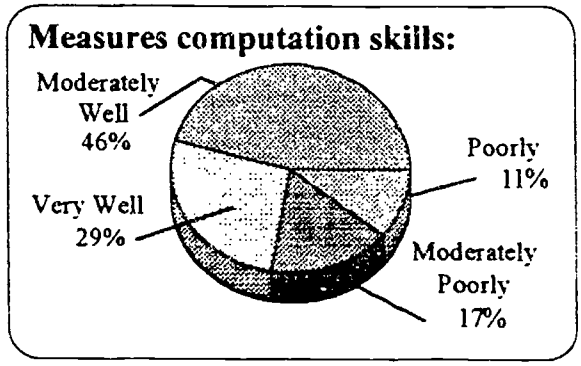
- If you didn't get one question, then you lose about 2 more.
- It included almost nothing of my past math experiences.
- Students get frustrated and nervous.
- On multiple-choice if you don't know the answer, you can at least guess.
- Long problems that are constantly the same can cause many mistakes. If you don't know one part about the answer then, most of them will be wrong.
- Because you have an answer on multiple choice and can see if they match yours.

- Make your instructions more understandable and have more variety to your problems.
- Thanx a lot and have a horrible day!
- If the task weren't quite so involved maybe it would have been better. Lighten up on it little.

Charts 37-39: Student Responses on the Survey:
How Well Did the Prompts Assess Mathematical Skills?
***Indiana Performance Assessments '92* (Mathematics-Communications)**







What did the teachers think of the math assessment?

Teachers' reactions to the assessments were cautiously positive and most of them would recommend replacing multiple-choice tests with this type. Almost all felt it measured computation, application, problem-solving, reasoning, procedural, and communication skills *well* or *moderately well*. But teachers acknowledged that many of their students did not like the experience. Perhaps this was, as some open-ended student responses

indicated, because the assessment was what a majority of the teachers saw as *somewhat* or *very different*. Over 90 percent of the teachers thought that the instructions to the teacher on the mathematics-communication assessment were either *very* or *somewhat helpful*; those to the student were rated a bit lower. About half of the teachers thought they were *appropriately challenging* and half *too difficult*. One out of five thought *Security One* was too easy. A majority thought too much time was allotted.

A complete table of the teacher responses to the survey about the mathematics-communication assessments is in the appendix, as is a list of the open-ended comments.

Summary: What can be concluded about the use of this type of performance assessment in Indiana?

1. How do Indiana students appear to perform on such a test?

Given the constraints under which some of the schools administered the assessments, student performance was acceptable. Examination of the overall results shows reliable increases in performance from Grade 9 through Grade 11, and there appears to be a highly significant increase beyond high school. With only a minor exception or two, this result is consistent across the six reading/writing prompts as well as in the math-communication assessments, where performance at all levels merits further study, analysis, and development. Generally speaking, the majority of students accomplished the task expected of them.

Not only has the project reported on the reading, writing, and computing abilities of Indiana high school students, but it has also answered key questions about the use of this type of holistically graded performance assessment for large-scale accountability and information gathering.

2. Can performance assessment be carried out successfully in Indiana?

From the outset there have been concerns that this type of assessment is too difficult to conduct and score—that it cannot be done. This relatively limited project has nonetheless successfully assessed over 5,000 students while conducting a survey and has proven that performance assessment is indeed feasible. In addition, it appears that it would be well-received by both teachers and students because it:

- Can be made more authentic to student needs and interests in terms of topics and tasks, thus making the applicability of the

skills and strategies assessed a clearer rationale for their being valued and emphasized;

- Reflects what we now believe about thinking and language behaviors by forcing the integration of thinking and computing with reading and writing rather than isolating such behaviors for an artificial instructional emphasis;
- Allows and even promotes individualized applications of the behaviors and knowledge assessed;
- Promotes the critical construction of meaning that depends on author purposes and a focus on audience;
- Integrates the knowledge of different subject areas in the realistic topics it uses;
- Creates a metacognitive emphasis in the application that promotes self-assessment and self-learning; and
- Focuses on behaviors, skills, and strategies that are now understood and valued in the community at large.

These are just some of the reasons that performance assessment answers to many of the key concerns about the extensive use of short-answer assessment in today's schools and its tendency to narrow and confine curricula and instruction.

The generally positive reception of the assessments designed for this project supports the contention that both students and educators understand, or at least sense experientially, the limited perspectives of much standardized testing and welcome an approach that assesses student achievement in an integrated, authentic way.

The study has revealed that there are a number of developmental issues, related, for example to the length of the testing times Many of the logistic concerns encountered in the trial can arise in *any* testing program using almost any instrument.

One of the more interesting aspects of the study has been that even as the teachers were aware that the math assessments developed and used were very difficult for their students, they yet endorsed them overall. Even their students, many of whom expressed a pronounced dislike of the assessments, appeared to recognize that they measure genuine computational abilities and that of communicating the problems and procedures of application.

3. Can such performance assessments be scored efficiently and reliably so that they win the trust of audiences interested in educational accountability?

The project did not include reporting the results of the study to the public, media, or others and registering their reactions to it; dissemination of the results to the schools did require the development of a carefully articulated reporting system. The study also included several analyses as to whether a scoring system can be taught to scorers, and a study of what is involved to make that training successful enough to produce reliable scoring. The development of scoring criteria as a rubric followed a painstaking and inductive procedure that rearticulated and verified the system through staff participation and two workshops.

It was found that the two-day training carried out in the project and following the rubrics and anchor example papers developed for each assessment can produce results that are at least 80 percent reliable for perfect correlation and that can be reliable within a scoring unit up to 100 percent of the time.

Given what was learned about scorer reactions to the rubrics and about the kind of more extensive training that scorers believe they need, it is concluded that the reliability of scoring by a carefully trained, statewide cadre is quite feasible.

While it did not formally study the reactions of the media and public to reports of the performance assessment scores, the staff did have numerous opportunities during the project to explain and present the idea to a variety of audiences, including educators, legislators, and businessmen. This response was more than accepting. The idea of evaluating student performance with an instrument that actually looks like a kind of reading/writing/thinking that a student will actually be expected to perform out in life was greeted with a grateful kind of enthusiasm.

4. How practical is the assessment approach?

Among the concerns about the feasibility of performance testing is one about how much time the approach requires. While it is readily admitted that the test requires two or three hours to administer fairly and that scoring cannot be accomplished by machines, the study has clearly indicated that the three-hour time limit is both feasible and possible to block out. There was indeed some expression of concern about the time limit being both too short and too long, but all students were able to respond to the tasks in some fashion meriting analysis.

Furthermore, it was discovered that scorers with just a half day's training are able to carefully score 15 papers each hour. This would mean that one scorer could conceivably handle over 100 high school responses in a single day. This study has indicated that performance assessment of the type tried out has advantages that other forms of assessment cannot have and that offset special considerations such as the need to develop a reliable scoring cadre. The conclusion is that this makes this form of performance assessment workable and reasonable enough to carefully consider.

Finally, there are subsidiary benefits to performance assessments. As the results of the summary data show, both students and teachers recognize

that this type of assessment required thinking and problem-solving. Students could not *get by* by merely selecting best guesses from the choices provided by the test developer. And, if teachers begin to teach to these new forms of assessment, Indiana students will be more critical thinkers and more able problem-solvers.