

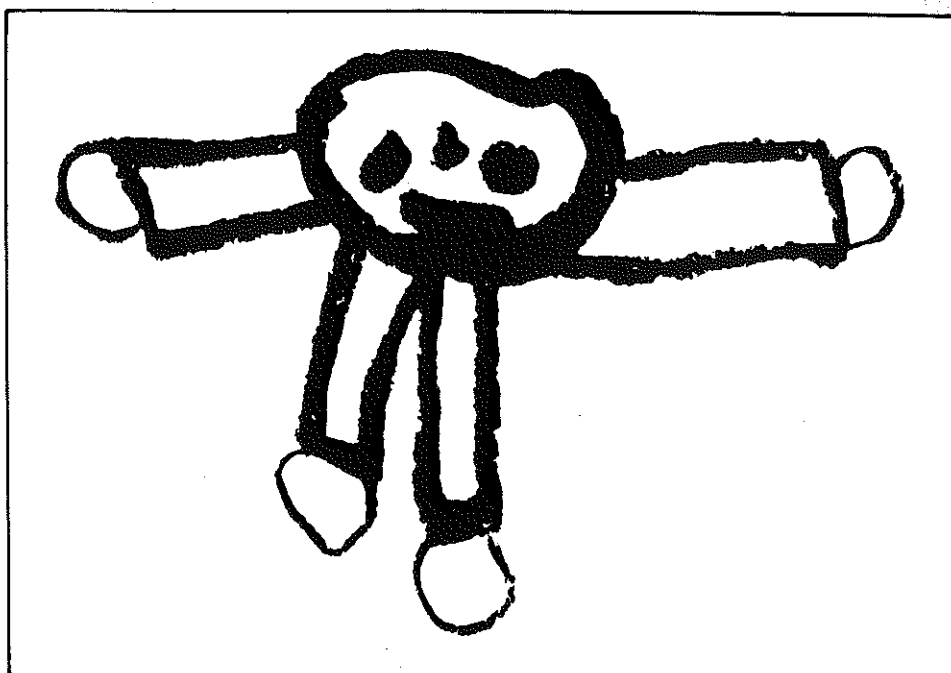
George Hein, Editor

**THE ASSESSMENT OF HANDS-ON
ELEMENTARY SCIENCE PROGRAMS**

In November 1972, educators from several parts of the United States met at the University of North Dakota to discuss some common concerns about the narrow accountability ethos that had begun to dominate schools and to share what many believed to be more sensible means of both documenting and assessing children's learning. Subsequent meetings, much sharing of evaluation information, and financial and moral support from the Rockefeller Brothers Fund have all contributed to keeping together what is now called the North Dakota Study Group on Evaluation. A major goal of the Study Group, beyond support for individual participants and programs, is to provide materials for teachers, parents, school administrators and governmental decision-makers (within State Education Agencies and the U.S. Office of Education) that might encourage re-examination of a range of evaluation issues and perspectives about schools and schooling.

Towards this end, the Study Group has initiated a continuing series of monographs, of which this paper is one. Over time, the series will include material on, among other things, children's thinking, children's language, teacher support systems, inservice training, the school's relationship to the larger community. The intent is that these papers be taken not as final statements—a new ideology, but as working papers, written by people who are acting on, not just thinking about, these problems, whose implications need an active and considered response.

Vito Perrone



George Hein, Editor

THE ASSESSMENT OF HANDS-ON ELEMENTARY SCIENCE PROGRAMS

Center for Teaching and Learning
University of North Dakota
August 1990

ACKNOWLEDGEMENTS

I gratefully acknowledge the generous support of the National Science Foundation (NSF), which funded the conference and the publication of this monograph through NSF Grant 3-2120. Special thanks go to Susan Snyder for her encouragement and support throughout the project. Sabra Price not only cheerfully carried out a myriad of everyday tasks associated with the conference and the monograph, but also commented on papers and made working on this project a pleasure. Laurie Beckelman provided wonderful editorial assistance. She gently but firmly managed to make each of us face the ambiguities of our sentences and the natural tendency to put things off. Finally, thanks to Emily Romney, as always, for her support, and to Jenny Hein for combing her complete collection of Peanuts books to find the cartoons that grace this volume.—G.E.H.

Contents

INTRODUCTION	
Assessing Assessment	1
<i>George E. Hein</i>	
PART ONE: Lessons from the Assessment of Reading and Writing	
Introduction	20
<i>George E. Hein</i>	
1 Lessons From Literacy	22
<i>Brenda S. Engel</i>	
2 Taking on Testing: Chapter Two	31
<i>Patricia L. Stock</i>	
PART TWO: Assessment Theory	
Introduction	66
<i>George E. Hein</i>	
3 Assessment and Teaching of Thinking Skills	68
<i>Audrey B. Champagne</i>	
4 Validity of Science Assessments	83
<i>Jerome Pine</i>	
5 Assessing Science Education: A Case for Multiple Perspectives	95
<i>Frank E. Davis</i>	
PART THREE: Large-Scale Assessments	
Introduction	114
<i>George E. Hein</i>	
6 What We Learn from State Assessments of Elementary School Science	116
<i>Joan Boykoff Baron</i>	
7 What Has Been Learnt about Assessment from the Work of the APU Science Project?	148
<i>Patricia Murphy</i>	
PART FOUR: Assessment in Science Education Research and Development	
Introduction	182
<i>George E. Hein</i>	
8 Assessment in the New NSF Elementary Science Curricula: An Emerging Role	184
<i>Maryellen Harmon and Jan Mokros</i>	
9 Assessing the Progress of Children's Understanding in Science: A Developmental Perspective	204
<i>Rosalind Driver</i>	

PART FIVE: New Approaches to Science Assessment	
Introduction	218
<i>George E. Hein</i>	
10 Young Children's Discussions of Science Topics	220
<i>Edward Chittenden</i>	
11 Children's Investigations of Natural Phenomena: A Source of Data for Assessment in Elementary School Science	248
<i>Hubert M. Dyasi</i>	
CONCLUSION	264
<i>George E. Hein</i>	
NOTES	280
BIBLIOGRAPHY	284

INTRODUCTION

Assessing Assessment

George E. Hein

ASSESSMENT, NOT TESTING

This book is about assessment, not testing. The distinction is crucial. In order to discuss assessment, we have to step back from considering only tests, the most common assessment form in education, and consider the full scope of the term. Assessments are judgments. In the most common use of the word, an assessment is a judgment of monetary value; property is assessed. In education, assessment refers to evaluation of educational outcomes. We cannot avoid the hard edge of its meaning: weighing the evidence, deciding, and sometimes finding a result wanting.

All attempts at assessment require definitions and raise powerful intellectual and emotional issues. What exactly is it that we are assessing? What are the criteria? What methods can we use? What evidence is available? Who is doing the assessment, and what are the qualifications of that person? What are the consequences of the assessment? Who gains or loses? The controversial and difficult issues raised by these questions in property assessment are evident to anyone who has been through the process; they apply equally to educational assessment.

TESTS

Formal tests are but one method for assessment, and a peculiarly American one. No other country, no other society, relies so heavily on multiple-choice tests, those familiar forms in which answers are marked by filling in circles with a soft pencil. The United States is unique in relying on this one kind of instrument to perform so vast an array of educational and professional gate-keeper functions. We know these tests so well, and they pervade our lives so completely, that we often mistake this one special form for the entire assessment enterprise.

Americans take over 100 million tests a year. These are used for a range of purposes, from determining school readiness, grade promotion, and awarding drivers' licenses to determining finalists for appointment as firefighters, real estate brokers, or insurance agents. Tests of this kind are so much a part of our culture that we can no longer determine how much they shape our

George E. Hein, Professor of Education at Lesley College, has recently completed a term as Dean of the Division for Advanced Graduate Study and Research. He is currently on a Fulbright Research Grant at King's College in London, examining data from the APU (Assessment of Performance Unit) surveys for information on children's acquisition of science process skills. He is a founding member of the North Dakota Study Group on Evaluation.

behavior as learners or teachers. American researchers recently went to Ireland, a country with a fine education system and a variety of assessment methods, so that they could measure the impact of standardized tests on an uncontaminated school system!

This vast use of tests does provide us with sufficient evidence to understand the limits of testing. No matter how thoughtfully developed, carefully administered, and cautiously applied, tests as a sole means of judgment have serious shortcomings. The insensitivity of standardized tests to special situations has been captured by one of our most famous cultural symbols: the Hollywood movie. In *Stand and Deliver*, a group of inner-city Hispanic youngsters are taught advance placement calculus by a quirky but dedicated teacher. They all pass the standardized test. But ETS refuses to accept their grades and assumes, incorrectly, that they must have cheated because a consistent pattern of wrong answers appears on the tests. Other stories of misassigned children, students who are inappropriately relegated to dead-end tracks, or inadequate professionals who passed qualification examinations abound.

The reason I emphasize the inadequacy of testing alone for making assessment-based judgments is that in our society we so frequently confuse testing and assessment. We treat the results of tests as if they provided adequate information to make sweeping judgments about academic, social, and political issues. The Coleman study, a major sociological research effort that examined the results of school desegregation, concluded that integration did not significantly help minority students' school performance. The researchers drew their conclusions primarily on the basis of an analysis of average test results alone. Despite the serious efforts of some of the study's critics, our society essentially ignored the interesting, aberrant scores from those predominantly black and integrated schools where minority children performed outstandingly. The secretary of education compares the educational systems of whole states on the basis of average SAT scores—this despite the fact that different percentages of students in the various states take the test! This results in the ludicrous conclusion that some states, which by any other standard do not have outstanding education systems, rank very high on the list. They appear to excel only because most students in the state don't even attempt the SAT. Some state departments of education are making judgments about the adequacy of teachers on the basis of students' test scores, with a difference of opinion about whether high-scoring or low-scoring school districts should receive additional resources. The assignment of children to special education classes solely on the basis of IQ test scores was common until recently, even though the

practice was seriously flawed. The policy was so entrenched court cases were necessary to stop it.

WHAT IS ASSESSMENT?

Assessment is a major and necessary component of education, as it is of many life activities. We want to know the results of our actions; we need to appraise the consequences of our work. In some fields, assessment is the essential, probably the only, important activity. In doing pure scientific research, assessment is the sole purpose of carrying out experiments: The *raison d'être* of science is to develop explanatory schemes. The reason to mix chemicals in a beaker, record data from the telescopic images, or administer the new medicine to some patients while giving others a placebo is to find out what happens!

Unlike pure science, most of life's activities have a purpose besides finding out the outcome of the action itself. We act so that students learn, roads are built, or food comes to the table. When we carry out these applied activities, assessment takes a less central role but is still crucial. The goal of education is to produce a change in the learners through the interaction between teacher and learners, to have the learners become knowledgeable in the areas taught. But all teachers also engage in the task of assessment. How do I know that the student has learned? What has the student learned? Is the student becoming competent and knowledgeable in the subject? These are not just idle questions, asked to satisfy the curiosity of teacher, administrator, or the general public; they represent an integral part of the teaching and learning process.

How do I know that I am an adequate gardener? I assess the results of my work. If the tomatoes constantly wilt, the morning glory seeds don't sprout, and the lawn repeatedly turns brown by the end of June (in the Northeast), my own confidence in my gardening ability would diminish. Probably the assessment of my gardening skills by others would be similarly diminished. If, however, my little urban plot is verdant, I pick ripe tomatoes by late July, and the morning glories and coreopsis provide brilliant color, both my own assessment and that of my neighbors will be quite different.

In any human activity, we are interested in the outcome, the result of our action. Although the assessments we perform may be rather informal in some situations, they are nevertheless part of the activity and they influence our future actions. I have learned through experience to plant primarily flowers that require little sun in my small urban plot in Cambridge, Massachusetts (I'm an expert on begonias, gloxinias, and impatiens), and I have modified

my actions on the basis of the results achieved. Similarly, all teachers, principals, and school system administrators assess the impact of their actions, and if they have any interest in their work, they are likely to pursue those practices that bring successful results. This is not the appropriate place to discuss varying definitions of success, but clearly, no matter what the definition might be, each person will assess and modify on the basis of the perceived results.

METHODS OF ASSESSMENT

What are the ways in which we can assess human actions? Whether we are talking about elementary science assessment or any other applied social science activity, we have only three major ways to find out something about humans: We can observe them, ask them, or note the results of their activity.

Behaviorist psychologists, who dominated American academic life for decades, refused to discuss such issues as intention, feelings, or mental models. Why children learned, how they felt about learning, or what might be going on inside their heads were considered beyond the reach of empirical study, and evaluators and researchers were urged to concentrate on what children did, on the relationship between the teachers' actions (stimulus) and the students' responses. Although most thoughtful commentators on education today no longer hold this limited view of the human experience, it remains inescapable that what is directly accessible to us in developing any assessment system, even in thinking about assessment, is behavior. We don't in fact know what children "learn," how they feel, and why they act as they do: We can make inferences about these attributes only on the basis of what we see children do, what they tell us, or what the products of their activity reveal to us.

Observation is by far the oldest, tried-and-true method of scientific inquiry: It was well established and accepted long before there was such a field as science. By the time Greek writers began to discuss science and describe rational inquiry into nature, observation had already established most of the regularities of the heavens, the basic knowledge of astronomy that was unchallenged until the invention of the telescope almost 2,000 years later. In our century, Piaget, in one of his early monographs, discussed observation as a wonderful method for finding out what children know, his only objection being that observation may be an inefficient way to assess something because the observer may have to wait for the desired behavior to occur. But direct observation of what children

do is still an important and powerful tool we can apply to assessment.

Observation is, of course, a scientific method applicable to both humans and nonhumans. Jane Goodall observed chimpanzees. Hundreds of others, from Aristotle to Dian Fossey, have observed animal behavior. Since we are interested in assessing human learning, we have another enormously revealing tool at our disposal, one based on the uniquely human ability to use language and, therefore, to introspect. We can ask people about anything. This leads to a host of assessment methods, ranging from informal conversations through formal interviews, to elicit verbal responses. A popular research tool for cognitive psychologists is to ask people to 'think aloud' while they perform some activity. Whether or not subjects' words accurately reflect immediate thought is open to debate, but this self-conscious language can be a rich source of insight into what people do when they solve problems, what they know, or what mental models they consciously draw on.

We can further exploit the human capacity for language by soliciting written information from learners. Diaries, questionnaires, lists, reflections, essays, and other written explanations all provide information useful for assessment.

Finally, many human actions lead to some result: They change the physical world. By looking at these changes, we can draw useful inferences. Archeologists commonly draw inferences about civilizations from the pottery and shards found at a dig. The kinds of fragments, the materials from which they are made, how they are disposed, and what they contain all provide information about how people lived. Likewise, the products of children's work—their drawings, the experiments they set up in a science lab, the way they use classroom materials, the materials they favor, their collections of insects or stamps—can provide valuable evidence useful for assessment.

Besides direct evidence from products, we can also obtain indirect evidence from the consequences of action in the world. The school janitor who finds lots of materials on the floor of the classroom might conclude that the children are sloppy, but he might also reflect that they have had the opportunity to interact with materials. The balances in need of repair suggest that they have been used, and science fair prizes awarded to a particular school argue strongly that something is going on in science in that school. A school system that provides science kits to classroom teachers may be able to estimate how much and what kind of science activity actually goes on from the amount and kind of wear and tear on the kits: If they are constantly returned in excellent

condition with all the components in place, they probably don't get heavy use.

All of these sources of information—observations, verbal reports, and products—can arise naturally or can be contrived for the purposes of assessment. Each tool can be sharpened and focused specifically for assessment. Thus, in a well-known study assessing the impact of an activity-based science program, Eleanor Duckworth (1978) provided children with a roomful of materials similar to, but different from, the ones the children had seen in the curriculum. She exposed children to these materials and observed and compared the behavior of children who had taken part in the program with that of children who had not. The experimental group carried out more activities and more diverse activities than the control group. This was a neat example of observation within a specially constructed assessment situation.

Similarly, we can overhear children's conversations or we can set out deliberately to talk with children, teachers, or others in the course of assessment. Both of these methods have been used. Paper-and-pencil tests, are, of course, a deliberate attempt to generate written information for assessment purposes. Likewise, the end-of-unit project that teachers review to determine what a child has learned uses the results of action as an assessment.

If we are serious in our effort to find out what children learn, how good our science programs are, what impact the money we spend on science has, or any of the many other assessment questions that arise, then we must take our task seriously. We must consider the wealth of assessment tools available to us, not just fall back on testing because it has been done before.

Each way of gathering information has strengths and weaknesses. Observation is powerful, but limited in both what can be observed and how often the relevant behaviors may happen; verbal reports always present the problem of sifting out what people say or write from what they mean. At best, verbal and written reports of behavior are only that: reports. They tell you only how people describe their behavior, which may have little in common with what they do. Consider that essentially everyone who drives in Massachusetts has passed a written multiple-choice test on the rules and regulations of the traffic code. You don't need much experience with Massachusetts drivers to know that their actual behavior differs from their assessment results.

In-depth interviews are a particularly powerful means for probing the meanings behind actions. That is why they are so popular with therapists and why Piaget adopted clinical interviews for his early work. But obtaining information this way may require

an enormous amount of time and resources, and the transcripts still must be analyzed to bring out meanings that may not be explicit. Products of work are also powerful tools (art and play therapists may prefer them to verbal responses as revealing inner meaning) but are often too complex for routine assessment purposes and may not be easy to summarize. Besides, not all kinds of learning lead easily to products. Tests, especially those with short answers, provide a certain simplicity and generalizability, and are relatively easy to administer, but they usually provide little insight into the reasons behind answers. Thus, the whole armamentarium of assessment should be considered in developing assessment systems.

THE CONFERENCE

I convened the conference that resulted in this volume because, along with many other science educators, I have great concerns about the current status of elementary science assessment. The last decade has provided us with a whole library of reports and articles critical of current educational practice and advocating reform. In the field of science education, critics are unanimous on at least one issue: Currently available science assessments are woefully inadequate. Whether we look at professional reviews of available tests published regularly in *Mental Measurements Yearbooks* or at a recent mammoth study of all elementary school tests from the Center for the Study of Evaluation at UCLA, whether we look at reports on student achievement in science and mathematics or reviews of the state of science education, whether we turn to reports about future funding priorities for the National Science Foundation (NSF) or the current state of research in science education, one message is always dominant: *We need new, more comprehensive, and more valid means to assess science learning.* Fortunately, a number of promising approaches may shed light on both the theory and practice of science assessment. These include the following:

- Both reading and writing teachers have come to realize that assessment may be conceived more broadly than testing and have begun to develop some interesting, comprehensive assessment schemes that reflect recent views of how children learn to read and write. These descriptions of the process of learning to read and write have much in common with modern notions of how children learn science, so assessment efforts in these curriculum areas may be of use to science educators.

- The last 20 years have witnessed the growth of a research industry devoted to understanding children's science concepts. Bibliographies with over a thousand articles on this subject have been collected; the third biennial international meeting on science misconceptions, held at Cornell University in 1987, attracted 352 participants and 162 reports (Novack, 1987). Researchers concerned with elucidating knowledge of science concepts use a range of assessment methods to find out what children know and believe. Since this research involves children learning science it is relevant to science assessment.
- A whole nation, the United Kingdom, has taken on the difficult and serious task of assessing science education through means other than standardized, multiple-choice, paper-and-pencil tasks, and some assessment experts in the United States have begun to learn from this work and apply it to our educational system.
- A recent wave of funding for elementary science curriculum materials by the National Science Foundation, the first in a generation, has provided both the opportunity and the need to assess the impact of these new curricula—initially on pilot groups of children and eventually on larger audiences. Curriculum developers do not use multiple-choice tests primarily, or even significantly, in the course of field-testing these materials.
- A small group of researchers has begun to explore alternative ways of assessing childrens' knowledge of science as part of an effort to connect assessment more closely with instruction.

I believed that by bringing together a group of people who could talk knowledgeably about these various perspectives, a group of teachers, administrators, state department science specialists, policy makers, curriculum developers and researchers, we could illuminate some of the issues, describe the state of the field, and make some informed statements about future possibilities.

With generous support from the National Science Foundation, I convened a conference of 30 people on a glorious, warm, and sunny weekend on Cape Cod in early November 1988. I asked nine people to write papers in preparation for the conference. These papers form the bulk of the material in this volume. At the meeting, we did not read the papers but discussed them, since they had been distributed earlier. Afterward, all the authors revised the

papers in light of the discussion, and two additional contributions on theoretical issues were solicited from participants.

The conference began with a discussion of assessment in reading and writing. Papers on the theoretical concerns that underlie any assessment followed, revealing some of the issues we must resolve before we can generate adequate ways to address assessment. Since we were concerned with alternatives to current practice, contributions from researchers who study children's science knowledge and understanding were useful. Thus, we included papers from authors who are primarily interested in research on children's science concepts and on curriculum development. Any effort to create science curriculum requires that the material be field-tested. How do we know that children learn from the material? The methods used by curriculum developers must be relevant assessments. Both at the level of large-scale assessment efforts and at the exploratory level of attempting to establish new practice, alternative assessment efforts exist. So this volume includes papers on some novel approaches used in state and national assessments. Finally, I solicited two papers on children's conversations and the products of their work as potential assessment tools to illustrate newer alternative methods.

This volume, as did the conference, owes much to the participants, all of whom contributed to the discussion. Several also provided prepared commentaries and chaired sessions. The general discussion has been a major source for this introductory chapter and the conclusion. As much as possible, I have tried to incorporate the views expressed by the various participants. But the responsibility for these chapters is mine, as I am also responsible for any omissions and reinterpretations of the views of others.

CONTEXT IN ASSESSMENT

One issue that appeared throughout the discussion at the conference and surfaces in a number of the papers is the significance of context in any assessment. At least half the papers deal with specific aspects of the contextual dimension of assessment. In various ways, they stress that the context in which assessment is carried out and interpreted is an integral component of the meaning of that assessment. The context of assessment influences the activity in a number of ways.

Context as Perspective

The context of assessment forms the perspective of collecting information. In considering the various ways in which we can

find out about children's behavior, we can see clearly that context is a powerful determinant of what we discover. If we want to observe what children do, the time and place of these observations is significant. We would have trouble finding out how proficient children are at manipulating the things of the world in a setting where these "things" were lacking, or where the children were inhibited from using them. If our method is to talk with children, then the nature of that conversation, whether oral or written, will profoundly affect the results. A whole generation of white, middle-class researchers decided that black children had "language deficits" based on formal verbal interactions between adult, white strangers and black children in school and clinical settings. The researchers who reached that conclusion appear never to have listened to the lively street talk that must have surrounded them on their way to the schools where they interviewed children. If we want to analyze the products of children's work for their abilities, then the result will be seriously influenced by the circumstances under which these products are produced.

These issues are particularly important under the conditions of contrived assessments. If we develop special situations, such as tests, to assess what children know, then we have to be particularly sensitive to the influence of the special situation on what will be produced. In a wonderful article about learning, David Hawkins (1974) points out that the specific learning behavior exhibited by rats in mazes can be obtained only if the rats are kept hungry enough so that food is a powerful enough attractor to motivate them to "learn" to run the maze. Well-fed rats exhibit much more exploratory behavior, wandering down dead-end corridors, and learning less quickly (or learning something else?) as they engage in a wider range of behaviors than simply finding the shortest path to the food. The image is useful when we try to compare the behavior that a test may solicit with the abilities that we actually want to assess.

The Context of Science

Context also is important in other ways. Scientists agree that science is a social activity, not a set of isolated learnings, and that it involves activities carried out over periods of time. Therefore, when we develop assessment instruments that measure what individual children know at any one moment, we may be missing major components of science knowledge. Again, the context of the assessment plays an important part.

Context in Interpretation

Scores on tests, results of judgments, or appraisals of knowledge have to be seen always within a framework of meaning and comparison. A major characteristic of most current state and national assessments is that they primarily present information on mean (average) scores derived from tests. This leaves out much of the meaning of the scores. For example, assume that we are interested in assessing the relative economic well-being of American families. We could do this by gathering data on *per capita* or family incomes in a number of communities. But without additional knowledge of context, this information can lead to serious misinterpretations. If we know that family A has an annual income that is 10 percent below the mean for its community, while family B has an income that is 10 percent above the mean for the same community, can we conclude that family B is better off than family A? Not at all. If family A consists of a retired couple with no mortgage and ample insurance, while family B consists of a young family with children, a large debt, and looming education costs, just the opposite may be true. The mean incomes for different communities also have varying significance, depending upon the economic conditions in that area, cost of living, etc. Common sense tells us that average statistical data need to be interpreted differently as we break down larger aggregates and apply the information to smaller units.

But this kind of oversimplification is exactly what happens repeatedly in the use of large-scale test results. Information that is of value only in a large context, say to determine the general level of knowledge or to compare one nation's scores with another, and which was gathered in a context that did not honor local differences, is then interpreted back to individual schools or communities (or even classrooms or individual children!) as normative about their achievement.

No matter how good a test may be, it should be moderated by an interpretation that takes into account the context of that score. As Brenda Engel points out in her contribution to this book, the "below mean" score on a test achieved by a child who has only recently learned English and has been making spectacular progress may require very different interpretation than a comparable score earned by another child.

We should also recognize that all assessment and testing occur within some context. It is a common misconception that standardized test scores are context independent, that they have

been freed of the necessity of interpretation from within a particular framework, and stand by themselves. This, of course, is far from the truth. In *Stand and Deliver* the “heavy” from ETS excuses his interpretation by stating that he is only a psychometrician and therefore free of racial prejudice. But the context of his belief is what leads him to the conclusion that a series of similar errors from one group of students implies cheating. In the absence of context, an equally plausible explanation is that similar errors from one isolated group of students derive from the students’ exposure to the same teacher. The role of the teacher is never considered by the psychometricians in the film. They emphasize repeatedly that the test interpretation involves only the students and ETS, not the teacher, his methods, or any other circumstances, except the ones they choose to consider.

One of the benefits of addressing the topic of assessment, rather than of testing, is that by thus viewing problems more broadly and asking which means are appropriate for eliciting the desired information, we are more likely to consider the crucial role of context. If we start with the premise that a short-answer test can tell us what we need to know about science achievement (and that we can legitimately decontextualize such a test), we will inevitably be disappointed when we try to generalize from the results. That particular method simply cannot carry the interpretive burden placed upon it.

However, if we start by asking the larger questions, and develop assessment instruments that take into account the context and what we really want to know, then our results may at least possibly satisfy our needs. The papers in this volume that are devoted to ways of finding out about children’s learning illustrate how researchers and curriculum developers do validate their work, and the discussions of state and national assessment efforts illustrate how performance tasks of various kinds come closer to meting out knowledge needs than do short-answer examinations.

RELIABILITY AND VALIDITY

Much of the conference discussion focused on two central issues of assessment: reliability and validity. If two different assessors were to carry out an assessment, no matter what the methods, how close would their results be? That is the central question of reliability. This seemingly simple question encompasses at least three different views, all of which were expressed at the conference.

What Do We Need to Do to Improve Reliability?

This question implies a technical approach. The issue is seen as similar to the problem of shooting arrows at a target. What can the archer do to make sure that arrows land relatively close to each other—not necessarily that they land on the bull's-eye, but that they consistently approach the same point? A similar task is familiar to anyone who engages in a craft, sport, or technical aspect of an art form. How can I serve the ball consistently into the service court, glide my bow smoothly across the violin's strings, or get my pie crust to turn a particular shade of brown? Framed in this manner, reliability, although not necessarily easy to achieve, is easy to conceptualize, and assessment experts can usually find a way to develop acceptable reliability no matter what the assessment task. When the writing community reestablished written samples as a means of measuring writing achievement, the assessment experts quickly developed a rating scheme to handle the thousands of short essays that must be judged. Olympic sports that rely on judges, as distinct from those that rely on direct competition between athletes, manage to achieve an acceptable system of agreement by providing several judges and averaging scores. The relative world-ranking of gymnasts and divers is not usually a matter of dispute.

Is Reliability a Reasonable Criterion?

One of the conference participants argued that too much emphasis on reliability may be unrealistic and an inappropriate pedagogic goal. Should all readers of an essay agree on its value? Such an outcome may be highly desirable for assessment purposes, but it does not mirror the world. Whenever I submit a piece of written work to a number of readers, I get back in return a great variation of views on the manuscript. If we put too much emphasis on reliability, do we distort our assessments?

What Is It That We Want to Obtain Agreement About?

Distinguishing between the assessment data itself and the meaning we attribute to it may be useful. Two well-informed assessors may agree on the assessment result (what the written answer to the problem actually consists of, what the drawing looks like, or what the words of the interview response were) but may differ on what that result means. It may be useful to separate our judgments about assessment results from the assessment itself. Professor Jim Miller of the University of Illinois (1988) carries out assessments of the general public's knowledge of science. He

concludes that this knowledge is shockingly low, based on answers to questions about whether the earth goes around the sun or vice versa, etc. Lauren Resnik (1987), on the other hand, argues that people may know more science than we think, but that they express their knowledge in everyday, rather than formal school, terms. I frequently find that students have difficulty explaining the difference between various ways of combining numbers, whether looking at averages and ignoring small differences is reasonable, or whether small differences are a significant component of the number set. Yet, most students understand why the following sets of results are interpreted differently:

Set I			Set II					
	<u>Grades on Tests</u>	<u>Term Grade</u>		<u>League Record</u>				
				<u>Win</u>	<u>Tie</u>	<u>Loss</u>	<u>Points</u>	<u>Standing</u>
Agnes	A A B B B B B B B C	B	Tigers	2	7	1	11	1st
John	A B B B B B B B C C	B	Lions	1	7	2	9	2nd

In short, separating the actual results of any assessment from the interpretation of those results is useful. The results may be highly reliable; the interpretation of those results may provide the opportunity for people to disagree significantly. The recent National Assessment of Educational Progress (NAEP) science achievement results (Mullins and Jenkins, 1988), which contained data on children's knowledge of science derived from multiple-choice, paper-and-pencil tests only, were widely interpreted as indicating the state of school children's science knowledge. NAEP spokespersons and others provided this interpretation, despite the fact that NAEP itself has published a booklet, *Learning by Doing* (1987), that argues in favor of performance testing in science, and despite the fact that both ETS (NAEP's parent organization) and NAEP staff have served on several panels that have argued for better assessment methods for science.

This discussion of reliability raises the parallel issue of validity. Reliability concerns the question of whether the same assessment results would be obtained under different conditions, at different times, or by different assessors. Validity concerns the issue of whether an assessment actually relates to the task we want to assess: Does elementary science assessment actually assess elementary science?

Again, the question is complex and can be viewed from both technical and social perspectives. Technically, validity refers not to how consistent the results are, but to how close they come to

the mark. If all our arrows fall in the same corner of the outermost white circle on the target, then our arrows consistently hit a mark, but not the one we want. If the ball consistently lands in the opposite back court, the bow always scratches across the violin strings, or the pie crust is uniformly black, we have achieved consistency (reliability) without achieving what we set out to do (validity).

A purely technical description of validity again leaves a complex but manageable problem, providing we can define what it is we are assessing. Even if I can't hit the bull's-eye, as long as I have a nice big target with a red bull's-eye I can easily determine how far I am from the center and can work to improve my aim. As long as I have the model of my mother's perfect pie crusts (even an imaginary model) clearly in my mind, I can tell how close I have come to it.

Unfortunately, assessing elementary science is a little like shooting at a target hidden by a blanket! We can tell when all the arrows cluster close together, but we don't really know how close we have come to the bull's-eye. We have not clearly defined it. A major issue in determining the validity of science assessment is to find some reasonable definition of what we mean by "elementary school science" or what we mean by a "good" or outstanding student-scientist, so we can validate our assessment scheme against a universally agreed upon principle. Traditional definitions of validity simply don't help with this problem. That the validity of our assessment appears 'on the face of it' to be appropriate by experts in the field, that the assessment scheme is consistent, or that it provides us with roughly the same ranking of students as other tests, are all inadequate for addressing the central validity issue: What is it we really want to assess? This topic is covered extensively in the paper by Jerry Pine, and tangentially in Frank Davis' contribution.

ASSESSMENTS AS INDICATORS

The discussion of validity leads to another issue that is relevant to all assessment systems. One way to be sure that an assessment is valid is to make the task itself the assessment. Thus, the practical part of the driving test, asking drivers to drive and park cars in traffic, has to be valid: The test consists of what a driver actually does. (Of course, this analogy reminds us of another problem associated with assessment, the context. All people who possess legitimate driver's licenses drove a car to an acceptable standard *under one set of conditions*. It doesn't take too many days on the road to realize that this standard is not met by all drivers at all times.)

To the extent that pieces of actual performance (if we can define what we mean by authentic scientific performance) can be made the assessment, we can be sure of the validity of our system. The best and most comprehensive assessment system would be a constant monitoring of all performance. Competitive chess and competitive tennis rankings achieve something close to this. The results of each match played are used constantly to recalculate the relative rankings of all players. A loss to a player seeded much higher is assessed differently from a loss to a player of equal rank, and a series of wins over players far below on the scale doesn't improve a player's ranking.

But obviously, such elaborate assessments, although appropriate to monitor the achievements of people in special high stress situations, are not practical in much of the real world. If we constantly used all the data available to assess everyone, we wouldn't have time left for anything else! And, although assessment is important for many activities in the world, it is not the only outcome of interest. Thus we will always struggle with the validity issue, as long as we separate some piece of behavior, or some less than total collection of results or products of work, and use these for assessment.

SAMPLING

The issue of sampling also appears repeatedly in many discussions of assessment. The American style of testing traditionally requires that every student take the tests: Whether we use the New York State Regents, the various commercial achievement tests, or the newer state mandated tests, most are administered to all students. Actually, for many assessment purposes, much smaller samples would be more than sufficient. This would save money and allow more time for instruction.

Americans take lots of tests, but they are subject to even more polls—market surveys, political opinions polls, television ratings sweeps, and many others. On the basis of polls that sample a small portion of the total population, we accept that candidates are front runners, have advanced or declined; that the population favors one policy or another; that it is knowledgeable about, or ignorant of, certain politicians' latest moral shortcomings. If the Democratic party wants to find out how well candidate X might fare against the incumbent Republican president, it can get a pretty good idea by asking fewer than 2,000 people how they would vote. Many national opinion polls use samples in the low thousands. Other widely reported polls survey only a few hundred respondents. But if care is taken to pick the right respondents, such small

samples are perfectly adequate to assure reliability and sampling validity. (That means only that the appropriate people were asked; whether the questions were valid questions is another, separate matter.) Yet, when states want to find out about school achievement, they often pick much larger samples of the student population, if not all of it.

Increasingly, state-level tests have moved to a type of sampling in which the whole population still participates in the assessment, but no one child takes the entire test. Some answer a few questions, others answer other questions, etc. In this way, assessors can gather information from an entire population without overburdening individuals within it.

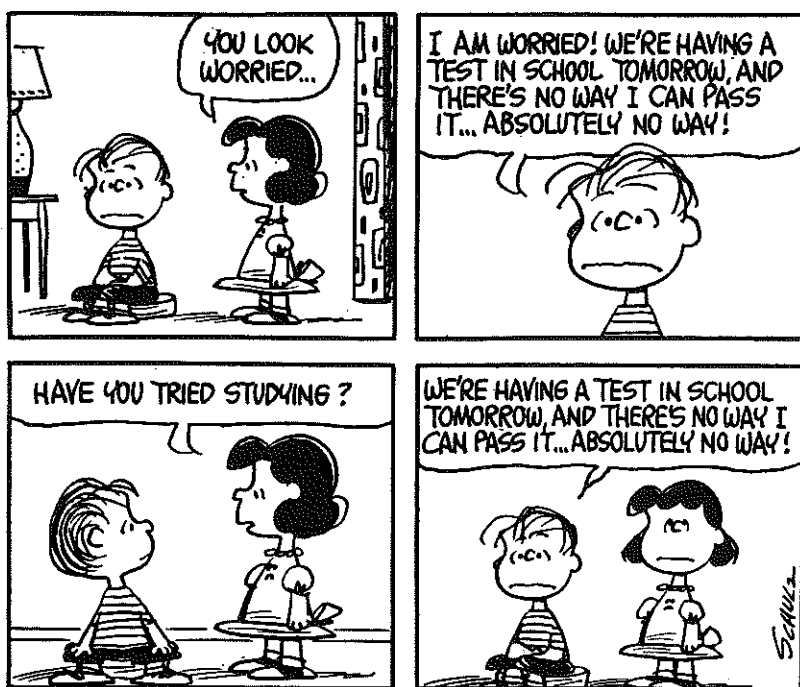
Educational assessment, unlike polling or other private surveys, such as marketing surveys, is carried out in a complex political and educational climate, the world of the public schools. Asking all the students to do something is often easier than trying to find a few selected students who would provide a representative sample of the population. One reason is logistical: The task of finding particular students or groups of students and making provisions for the assessment while simultaneously monitoring the rest of the school population is enormous. Another reason is that when all students are tested (or assessed), they are more likely to take the task seriously. It fits into the general mentality and experience of what school is like. If a few are selected to participate in the assessment, all sorts of questions have to be answered, and individuals may attribute being chosen to a wide set of reasons.

CONCLUSION

The papers that follow elaborate and expand on the themes I have raised here. Each discusses assessment from a different viewpoint; each covers a unique component of this complex subject. Yet there are themes which repeat from chapter to chapter. Viewed from the broad perspective of assessment, rather than through the narrow lens of testing, surprising commonalities emerge. I hope these will become apparent to the reader. In the final chapter, I will address these topics.

PART ONE

Lessons from the Assessment of Reading and Writing



Introduction

George E. Hein

Science is only a minor component of the curriculum in most U.S. schools, and it only recently has begun to receive “serious” attention. The new state-mandated achievement tests include sections on science, as do newly mandated state curricula. This development brings into sharp relief the relationship between instruction and assessment, and the extent to which tests determine curriculum. This relationship has been apparent for years in the traditional school subjects, especially in the language arts, which hold the predominant position in elementary instruction.

Concerned teachers of reading and writing have long struggled with the dichotomy between traditional testing and conceptions of teaching and learning. English teachers, through the National Association of Teachers of English and other organizations, were at the forefront of the struggle to reestablish writing as a component of writing assessment—at a time when test makers were arguing that writing could be assessed through short-answer questions alone! Reading specialists interested in whole language instruction have developed a sophisticated set of arguments and assessment tools to support their positions in favor of alternative assessment strategies.

The field of science education has much to learn from alternative assessment work developed during the last decade in reading and writing instruction. The following two papers present alternatives in each of these curriculum areas. The papers complement each other. But they also provide differing perspectives.

Both authors argue that assessment congruent with developmental and interactionist theories of learning must provide means for both learners and teachers to participate actively in the assessment. They describe assessment methods that result in qualitative evaluations, collections of documents, and recognition of the role of context in any assessment scheme. Both also discuss and illustrate the relationship between assessment and instruction, focusing particularly on that component of assessment most useful for teachers. Brenda Engel provides a general description of recent work in reading assessment; Patricia Stock focuses on a specific example of writing assessment.

The two papers also suggest fascinating possibilities for science instruction. What are the preschool science activities that parents can and should encourage—activities that parallel the early reading and language experiences so important for children? Can we build a ‘library’ of science activities to match the library of books that are available to complement and reinforce children’s individual interests? What would a discussion of children’s science that parallels the discussion of a sample of writing be like, and how can such discussions be initiated and legitimized in schools?

Lessons From Literacy

Brenda S. Engel

INTRODUCTION

Although science and literacy are very different fields, they have two things in common: Both are basic areas in the school curriculum and both currently offer new opportunities for educational theorists and evaluators. This paper assumes that recent international research on literacy learning can yield lessons for the evaluation of science education. It explains the theory resulting from this research, examines the effects of traditional evaluation on the twin issues of pedagogy and evaluation in literacy learning, considers some alternative methods of evaluation, and, finally, discusses possible implications of the foregoing for practical change.

First, however, I want to consider the relationship between teaching and evaluation. Despite our knowledge to the contrary, we act as though we teach subject content, then test students to see what they have learned. In reality, the relationship between teaching and evaluation is interactive: Evaluation influences what is taught, how it is taught, and when it is taught. It acts as both a constraining and determining influence on the curriculum.

In the fields of both literacy and science, evaluation is to some degree responsible for education getting stuck in an outdated paradigm. It's risky for teachers to be creative with new subject matter or teaching methods when the outcomes of their teaching will be assessed in ways that remain static and inflexible. Because they are quite reasonably reluctant to expose themselves to accusations of incompetence or their students to the possibility of failure, most teachers avoid taking risks and, instead, "teach to the test."

If we recognize that evaluation interacts with teaching, however, evaluation takes on a different function and will itself need reevaluation to go along with changes in educational theory and practice.

LITERACY LEARNING DEFINED

Psycholinguists have come to understand literacy learning as a process by which the individual constructs knowledge rather than as a process by which the individual acquires specific skills and techniques. Knowledge, to be broadly useful, must be built anew by each individual learner. Lessons "received" without being

Brenda S. Engel is an Associate Professor in the Lesley College Graduate School, Division of Advanced Graduate Study and Research. She has published articles and monographs on qualitative evaluation methods and, for the past six years, has been closely associated with the Cambridge-Lesley Literacy Project.

constructed—that is, without being fully integrated into a person's existing understanding, experience, and feelings—remain thin, lack flexibility and opportunity for self-improvement, and, perhaps most important, allow no place for invention, playfulness, and pleasure.

Don Holdaway and other educational theorists (Holdaway, 1979) have elucidated the processes by which the individual constructs knowledge of speaking, listening, reading, and writing. They emphasize that all language learning is developmental. Infants, from their first day, begin to learn language—first responding to the tones and rhythm of the human voice, later volunteering their own baby noises, recognizing the meaning of spoken words and sentences, and soon beginning to approximate actual speech. Baby talk, which parents and other family members have an uncanny way of understanding, gradually evolves into language that even strangers can understand.

Children also simulate reading, a stage Holdaway calls *emergent reading*. A child will paraphrase the words in a familiar book while turning the pages, sometimes putting in a few recalled words or phrases or even reciting the entire memorized book so that a casual observer finds it hard to believe the child is not actually reading. Exaggerated expressiveness, imitative of adults reading aloud, often accompanies this kind of simulation. In the natural course of events, emergent reading turns into early reading and eventually into fluent reading. Similarly, young children often go through a stage of pretend writing, simulating lists, menus, notes, signs. First, letter-like shapes appear, then recognizable letters, and, finally, readable words and sentences.

Desire to participate in the world and to gain control over the facilitating elements of the culture, one of the most important of which is clearly language, drives the infant's movement toward language competence. Motivation, if the term is even appropriate here, is intrinsic, and learning, consequently, becomes a "self-improving system" (Clay, 1979).

Researchers, including Jerome Harste and his colleagues at the University of Indiana and Yetta and Kenneth Goodman at the University of Arizona (Goodman, et al., 1984; Harste, et al., 1984), have demonstrated that children, before going to school, know a lot about both the conventions and the forms of literacy. They can "read" street signs and names on cereal boxes; they are familiar with a variety of logos and often can recognize their own names. What's more, as Ferreiro and Teberosky point out, this kind of early knowledge is not confined to children in advanced industrial societies (Ferreiro, Teberosky, 1982). Young children, probably the world over, as they become aware of the meaning of

everyday symbols in their culture, become emergent readers and writers.

According to recent research, language learning is not only developmental but also holistic: a continuous, natural process occurring in a context of meaning. As a father brings a spoon to his six-month-old's mouth, he says, "Open wide!" A mother warns her two-year-old that the stove is "HOT." Even the bedtime story is read aloud in a context of meaning—pleasure in the quality of the book and shared experience. Meaning is a matrix that holds the pieces in place as the child learns. Ethan at two-and-a-half can name the members of the Celtics starting team; knows the difference between a free shot, lay-up, and stuff shot; and can sit still, watching a basketball game on TV, for periods of up to an hour. Although he has little if any concept of the structure of the game and would be unable to talk *about* it, he senses its meaning through the responses of his parents and their friends—their enthusiasm when the home team makes a basket, tension at a close score, and, finally, relaxation—either happy or disappointed—when the game is over.¹

Early learning experiences have other notable characteristics. Each step forward the infant makes is recognized and celebrated—the first smile, word, sentence. The child is encouraged and rewarded by the delight and recognition of adults. Parents or caretakers also keep track informally of change and progress. They are likely to know, at any point, what ideas the child will understand, what new experience or materials to offer. In addition to serendipitous experiences (such as watching basketball on television), more intentionally educational activities such as reading a storybook, taking a trip to the zoo, or giving the child beads to string are planned according to ongoing evaluation by caretaking adults.

EVALUATION

Early learning can be characterized as developmental, holistic, and driven by meaning; the young child's motivation is intrinsic and natural. The child moves from "clumsy approximation" (Holdaway, 1979) of adult performance to competence. Evaluation of learning, before school, takes the form of positive, essentially descriptive, "keeping track."

What, then, happens when the child enters school? The continuum of learning is interrupted, in part because schools generally fail to recognize the knowledge children bring with them—fail, that is, to see school as an extension rather than an initiation of literacy. Schools assume that all literacy learning begins on the first day of formal schooling, a form of denial that

often makes it difficult for children to connect their previous knowledge with school learning or to feel that it is valued. Consequently they don't see the positive role reading and writing can play in their lives. Those children familiar with many printed symbols but who have not had experience with *literature*—with being read to at home or in preschool—are particularly vulnerable. They often have a hard time finding meaning or interest in phonics or decontextualized sight words, which they may construe as “things to be learned in school,” rather than as elements of an activity important to their lives.

Compounding this problem is the assumption, inherent in traditional pedagogy, that teachers possess knowledge that they transmit to children: Teachers teach; children learn. Children are viewed not as active seekers, but as receivers of already constructed knowledge. Their early impulse toward learning and using language in all its forms is interrupted as initiative shifts from child to teacher. As a consequence, intrinsic motivation, which requires that the learner perceive some personal relevance or meaning in an activity, is sacrificed. When given at all, the reasons for undertaking the often painful process of learning to read and write are likely to be couched in terms of standard expectations, passing tests, and future rewards.

Grades, the threat of failure on tests, and the encouragement of competition among children provide substitute motivation; grade-level achievement based on national norms, rather than development or movement toward competence and control, constitutes the basic standard.

Despite our national rhetoric about recognizing individual rates and styles of learning, the value of progress in itself is rarely validated. Norms overwhelm individual histories of accomplishment even when the latter are impressive. Thus an ESL student who has made dramatic progress in three years of schooling, reading and writing in two languages, fails the state basic skills test in grade 3 because she is not “up to grade level” in English. Another child who entered school with less than the usual knowledge of the conventions of print but is rapidly closing the gap between himself and his classmates is also judged inadequate because his “skills” are still below standard expectations for his age/grade. Early failures, of course, are hard to undo; they tend to govern a child's sense of self as learner.

What about the qualities of holism and meaning that characterize early learning? Once a child enters school, even kindergarten, in most cases recognition of meaning is no longer considered a necessary condition for learning. Fragmented skills—long vowel sounds, terminal consonants, word definitions, and the

like—are both taught and tested outside of meaningful context. The purpose of such activities is not evident to many children, and children's enjoyment is not valued.

Finally, the positive "keeping track" of learning that occurs at home gives way to judgment in the schools. Instead of describing where the child is "up to," national and local standards prescribe where he or she *should* be. Prescription replaces description. Standards in general become externalized and no longer relate to individual history and development. By "objective" standards, children are too often found wanting.²

These failures at a young age rarely work in the child's favor. An early sense of inadequacy can become a self-fulfilling prophecy. For instance, research indicates that high school students held back in the elementary grades are prime candidates for dropping out of school later.

Evaluation, interactive with pedagogy, is implicated in all these losses: Tests, which are judgmental by their nature, reward received rather than constructed knowledge. The essential characteristics of child-initiated learning—development and meaning—no longer have presence or authority. Motivation, once natural and intrinsic, becomes contrived and extrinsic. Pleasure has no place at all.

ALTERNATIVE EVALUATIONS

Although standardized testing in the field of early literacy has lost little of its actual hegemony, it has been strongly criticized recently, particularly by early childhood educators. Alternatives are being suggested.

In order to conceptualize radically different ways of evaluating learning, it is important to keep in mind the long-range purpose of the activity under scrutiny. The purpose of learning to read, for example, is to read—to construct meaning from print—in order to make accessible an extended world of knowledge, information, and experience. This may seem obvious, but in schools the purpose of learning to read is daily stated as "raising test scores." The difference is important. If the purpose of learning to read is reading, then evaluation should be directed at reading, not at fragmented knowledge of phonics, decontextualized word meanings, or other piecemeal skills which, added together, do not constitute reading. Many children who have learned successfully to decode, for instance, fail to take in the meaning of what they have read.³ The main reason fragmented skills don't add up is the lack of context, the larger understanding within which a text has meaning.

Context is integral to reading and must be taken into account in both teaching and evaluation. The context can be interest in and knowledge of a subject—sharks, for example—or familiarity with a form, such as fairy tales, but without a grasp of context, the reader flounders; the words lack significance. Only once the context is established can the reader make sense of the text. We've all experienced reading about an unfamiliar subject and finding that the words, though familiar, don't come together in ways that signify. Recognition of the value of the activity as an integrated whole is basic to the following suggestions for evaluating literacy learning.

Keeping track of children's progress into literacy can begin, for emergent readers, in preschool or kindergarten, through a *literacy interview*: The teacher/evaluator invites the child to choose a favorite book to "read" with her. With the tape recorder on, the child and teacher together explore aspects of literacy, as the teacher reads the book: "Can you show me where the front cover is?" "Point to where I start reading. And then where do I go?" "Can you show me one letter? One word?" "Do you know any of the letters on this page? Any words?" "What letters are in your name?" "Do you know what this dot [period] means?"

The teacher encourages the child to paraphrase the story and to anticipate what will happen next and how the story will end. He or she may ask related questions, too: How much TV does the child watch at home? Is the child read to at home? What kinds of books does he or she prefer? In other words, the teacher constructs a literacy profile of the child that takes into account previous experience and knowledge. The profile can be summarized in the form of an "inventory" of early literacy characteristics. These characteristics must be strictly distinguished from objectives or checklists, however. They describe rather than prescribe.

Recorded interviews and inventories provide baseline data for keeping track of early literacy development. Later, as children become early readers, the classroom teacher can make periodic tape recordings of oral reading samples, using graded texts. Basal readers, which are not always adequate as the main fare of a reading program, are useful for this purpose—marking progress systematically.

Another useful way to evaluate literacy learning is *miscue analysis*. Since Kenneth Goodman originally elaborated the concept of miscues in 1971, a number of schemes for miscue analysis have been developed and published in New Zealand, Australia, Canada, and the United States. A miscue, according to Goodman, is "an actual observed response in oral reading which does not match the expected response" (Goodman, 1973). The

significant factor in miscue analysis is the differentiation of meaningful miscues from meaningless ones. A *meaningful* miscue does not destroy the sense of the sentence: "They built themselves a comfortable *home*," for instance, instead of *house* as printed. A meaningless miscue occurs in the sentence, "They built themselves a comfortable *horse*." Omissions, substitutions, and insertions that don't change the sense of the sentence are counted as "meaningful miscues," common to and excusable in all good readers. The purpose of miscue analysis is to determine if a child is trying to make sense of a text rather than simply decoding it.

Another significant indicator in miscue analysis is the ratio of self-corrections to the total number of miscues. This ratio gives an idea of the child as autonomous reader, taking on some of the functions ordinarily assumed by the teacher. When children begin to correct their own errors, they are developing "self-improving systems" (Clay, 1985), an important goal for all reading programs.

The recent, wide-spread process writing movement has brought writing across the curriculum into equal partnership with reading. As a school subject, writing has traditionally been assessed in terms of children's control over conventions such as spelling, handwriting, punctuation, and grammar. Again, however, if we consider the long-range purposes of writing—to construct and communicate meaning—we are challenged to develop new ways of evaluating children's written work as well. Increasing control over expressive language, not merely over conventions, characterizes the developing writer. Although writing can be described to some extent by word and sentence counts, describing actual quality is a more difficult task, as yet unsolved. Keeping periodic samples of children's work, however, is an effective, nonjudgmental way of keeping track. The work itself provides concrete, persuasive evidence of change and development.

This brings me to a final, comprehensive suggestion for evaluating literacy learning: ongoing collections of children's work—writings, drawings, and reading tapes as well as teachers' notes and observations, lists of books read, interviews, and inventories of various kinds. This relatively uninterpreted documentation or portfolio is interesting, authoritative, and immediately useful as feedback for teaching. It also avoids most of the negative effects of evaluation described earlier in this essay. Its main disadvantage is lack of formulation and concision, both necessary for accountability to administrators and the larger public.

USING ALTERNATIVES

All of the alternative forms of evaluation I've discussed assume that teaching and evaluation are interactive and therefore involve teachers in a new role. Teachers must play a more active role in evaluation, which itself must be reconceived in terms of differentiated audiences. Information useful as feedback for teaching is primary evidence: ongoing documentation of children's work. It needs to be translated into a concise, probably quantitative, form useful to administrators and policy makers. How this is done is beyond the scope of this paper; it is a question of summarizing qualitative data—not an insuperable problem. I will suggest, however, that each child's cumulative work folder can be reviewed annually and evaluated for development of initiative, independence of thought, imagination, breadth and depth of knowledge, as well as control over conventions. Teachers and evaluators together can develop a consistent format to represent the results of this review—inventory, anecdotal comment, a checklist of recommendations, or whatever method of summing up seems most appropriate within the particular educational context.

Some of the positive conditions that description (as opposed to judgment) enables pertain not only to literacy learning but to all education. They include emphasis on development, holism, and meaning. Therefore, if we recognize education and evaluation as dual aspects of the same endeavor, the following principles hold true:

- Primary responsibility for evaluation belongs to teachers with evaluation seen as feedback for teaching.
- Evaluative criteria should be viewed longitudinally and related to the history and character of the individual rather than to national norms.
- We must recognize pre- and outside-of-school knowledge.
- Subject matter should be kept intact, with skills taught in context.
- Intrinsic motivation and children's impulse toward self-correction should be recognized, valued, and used.
- Knowledge should be seen as constructed by the individual within a learning community and as developmental in character.
- All evaluation should be based, as far as possible, on real and purposeful activities.

What, then, are the possibilities for changes in the practice of educational evaluation? Although standardized testing still dominates educational evaluation, a number of states and communities, dissatisfied with standardized tests for a variety of reasons, are exploring alternatives. These alternatives include changes in the official definition of reading, suspension of early testing, development of observational instruments, and specifications for student portfolios, to mention a few.

The time is ripe to formulate *adequate* alternatives to traditional evaluation, alternatives that will be both more informative and more conducive to good education. These alternatives must also provide methods of accountability to the educational hierarchy, parents, and the tax-paying public. Until we have alternatives that meet all of these requirements, the status quo will prevail.

A period of overlap may be necessary during which we try out a new and more rational system of evaluation based on observation and organized documentation. If the new system successfully provides *more* and more *useful* information, has a positive influence on educational methods, and can be sold to parents (probably the biggest hurdle), the old ways may begin to lose power and eventually atrophy. That's the hope. Alternative methods will probably have a better chance of taking hold in the kindergarten and primary grades, where standardized tests visibly do more damage (NASSB, 1988) and where competition for higher education is not yet a pressing matter.

The criteria for evaluating evaluation must be related to both purpose and effect. The purpose of evaluation must be to encourage the education of the individual and to make learning *more*, not less, accessible; the effect has to be the same. Although this paper will not specify exactly how the foregoing discussion of literacy evaluation might be applied to evaluation of science education, it can be stated, as a general principle, that any evaluation that unreasonably constrains teaching and discourages the individual learner cannot be justified—whether in the name of policy-setting, international competition, or “excellence.” Evaluation must first and always serve children who begin life eager to learn. It is that eagerness that we, as educators, have a responsibility to nurture, not destroy.

Taking on Testing: Chapter Two

Patricia L. Stock

Good writing teachers typically mistrust most writing assessments, and for very good reason. As writing assessment is now customarily conducted, conflict exists between how writing is tested and how writing is taught in classrooms where students—in their own and their teachers' eyes—seem to be learning to write purposively, meaningfully, and effectively. In the opinion of many teachers, testers rarely test what good writing teachers believe they should teach; in the opinion of student writers who are making progress and taking pleasure in doing so, testers rarely provide them opportunity to compose in written language for purposes meaningful either to them or to the audiences for whom they typically write. And in the joint opinion of good writing teachers and students who are learning, testers—for reasons that are understandable, given present conceptions of their task—focus attention upon features of tests and factors in their production that, all too often, seem marginal rather than central to meaningful characterizations of something that might be called "good writing."

Stock and Robinson (1987, p. 93)

Since the summer of 1984, with colleagues at The University of Michigan and in the public schools of the cities of Ann Arbor and Saginaw, Michigan, I have been working to develop local assessments of students' writing competencies. My colleagues and I have been arguing, on both practical and theoretical grounds, for the replacement of the large-scale assessment of students' writing—as it is customarily conducted today—with local, teacher-developed assessments of students' writing (Barritt, et al., 1986; Clark, 1983; Stock and Robinson, 1987)¹. We have offered the following practical argument for local assessments of students' writing: Assessments are more often designed to serve the needs of administrators and policy makers than teachers and students. Too often, these constituencies see themselves as having different needs. As they see their needs, administrators must demonstrate to policy makers, who, in turn, must demonstrate to the public, that learning is happening. For this purpose, the broad generalizations and numerical summaries of the sort that large-scale assessments are designed to produce are viewed as generally acceptable vehicles for conveying such persuasive information.

Patricia Stock has been a teacher of English and composition at the high school and college levels for almost 20 years. Working primarily in recent years as a teacher of teachers, she is presently writing a book and articles about the politics of teaching and assessing literacy.

These generalizations and summaries also serve the needs of administrators and policymakers who look to the findings of large-scale testing agencies to guide them as they develop curricula to meet students' needs. However, because large-scale assessments offer gross findings only (as those who conduct them acknowledge), teachers and testers alike recognize that they cannot help teachers who would teach better by responding to the particular needs of each student. For this reason, large-scale assessments are not appropriate vehicles for helping teachers learn what they need to know if they are to teach their students to write more effectively.

The argument based in theory that my colleagues and I have offered for local assessments of writing has been this one: The practices characteristic of large-scale testing are rooted in two necessary assumptions that conflict with current theories of language learning and language use. First, since large-scale testing regards students' texts as artifacts, as products that can be analyzed, such testing assumes that texts can be isolated from both the human intentions and actions involved in their production and from the contexts in which those intentions and actions are realized; and second, since all large-scale testing relies upon criteria that are pre-set and uniform for all administrations of a test, such testing assumes that individual essays students produce in concrete and particular circumstances can be appropriately evaluated with reference to criteria that are insensitive to the constraints that local circumstances impose upon writers and their writings. Because large-scale assessment is atheoretical at its base, it cannot lead to characterizations of students' writing abilities and achievements that administrators and policymakers need in order to make sound educational decisions any more than it can supply teachers with the information they need to improve their teaching of writing.

In this essay, referring to work undertaken by several of us at The University of Michigan together with practicing teachers and administrators in the Public Schools of the City of Saginaw, I propose to demonstrate how the assessment of students' writing, when undertaken by those students' teachers, can enable teachers to learn what they need to know to become more effective teachers of their students.

THE SAGINAW TEACHERS' ASSESSMENT OF WRITING

Those of us who have worked together since January 1986 in Saginaw to develop an assessment of students' writing have done so within local circumstances. Saginaw is a middle-sized, largely industrial city that has suffered many of the afflictions

visited upon cities in the Rust-belt whose economies have depended upon ailing industrial corporations. Its schools have been consistently, if not always adequately, supported financially (no millage has failed in over ten years, not even in the time of the recent severe recession); committed and effective teachers may be found in all buildings; imaginative administrators have found both ways and means to support teachers and us in our work with them and in our joint work with students. We have also worked together within the peculiar sets of political and pedagogical constraints that are imposed by the larger contexts of schooling in the United States: the public's and politicians' faith in standardized testing and in its results; state and national proposals for school improvement whose mandates are insensitive to local needs and possibilities; inequalities in purportedly equalized public funding; the politics and publicity of alleged school failure—e.g., the Nation at Risk syndrome that has been exploited by politicians, and especially by those politicians who see schools merely as the servants of corporations and of the corporate state. And we have worked together within other constraints that inevitably arise in multi-racial and multi-ethnic communities (53 percent of Saginaw's students are black, 32 percent white, 13 percent Hispanic).

As we have worked, those of us who are university teachers have had to establish new relationships with our colleagues who teach in schools in order to enable free and critical dialogue. Because it is my opinion that in our work we have become co-equal teacher-researchers working together to find solutions to problems we hold in common, I have chosen to speak in a common voice for all of us as I describe the work we did together.² As teachers working together, we have had to establish different relationships with administrators and school board members to alter well-established structures of authority and equally well-established patterns of action. And new relationships have had to be formed among those of us who are black and those who are white and those who are Hispanic, if for no other reason than to find common spaces in which we might negotiate over the meanings that attach themselves to language differences in communities that are socially separated along racial and ethnic lines.

Sensitive to the fact that we would have to find our way as a result of our collaborative efforts, we began our work guided by one certain imperative: The school teachers among us insisted that the one day every other week they were taking away from their students must not only lead to a writing assessment eventually but must also profit their classroom teaching immediately. This requirement, that theory and research inform practice and that they

be made subject to testing through practice, shaped the content and the conduct of The Saginaw Teachers' Writing Assessment.

From the outset, we decided that the strategies we would use to conduct our research would be strategies we could take back to our classrooms and teach our students to use in the service of their learning. We also decided initially to divide ourselves into two study groups, each of which would investigate issues we hoped would help us improve our practice as teachers of writing even as our inquiry into these issues taught us lessons we needed to learn if we were to assess students' writing meaningfully. Members of one study group began investigating both the "school" writing we were requiring of our students and our practices for evaluating that writing; members of the other initiated letter writing exchanges among their students in an effort to determine if students composed more effectively when writing to their peers than they did when writing for their teachers.

The School Writing Group

The "school writing" group took as its central tasks to describe the kinds of writing teachers typically ask of students and to make explicit the values for language use that were guiding us as we evaluated students' writing. Having invited each other to bring to our meetings sample assignments and samples of students' responses to those assignments, we had ample material upon which to base our inquiry. These assignments and papers always generated engaged and sometimes even heated discussion about the different values different teachers were bringing to bear upon students' writing.

In the very first meeting at which we read and commented on students' papers, some of our different values became apparent. On this occasion, our method of inquiry was to read one student paper at a time and to have every member of the group write for themselves protocols of their readings. Once teacher-readers had "written their reading," we shared our readings with one another. In his essay *Literacy and Conversation: Notes Toward a Constitutive Rhetoric* (Robinson, to be published), Jay Robinson captures the purpose and spirit of the group's collaborative inquiry. Reflecting on one of the essays we read during that first session and on a variety of the readings of it that we recorded as a research group, Robinson names that variety and demonstrates that while an adequate reading of a student's text can emerge from the interplay of multiple perspectives, an informed reading can emerge only when that reading takes place within an interpretative community.

Entitling it *Learning From Experience*, Robinson awards the student writer a pseudonym, Fred Albright, and reproduces Albright's impromptu response to "a bit of gnomic wisdom": "The wise man learns from others' experience; the fool learns from his own."

Learning from Experience

I think that the statement that "the wise learn through the experience of others, fools learn through their own experiences" means that a person who is wise see's someone else's experience and if its a bad experience for that person, the wise person will try not to have that same experience. Well the fool won't pay any attention to that and he might have the same experience as that person, and if he tried to avoid he might not of been hurt.

I think a example of this is getting pragnent befor you are married. Say Mary had a baby, she has a wise friend named Sue, and a foolish friend named Amy. Mary goes through alot of problems like desiding if she is to have a abortion put it up for adoption, keep the baby, when to tell her parents, etc. Sue see's her having these problems, so when her boy friends want's to have sex she says no. But Amy say's it wont happen to her and she has a active sex life and she finally get's pragnent and she has to go through the same things that Mary went through because she didn't pay attention to Mary's experience.

I didn't write that example to say you shouldn't have sex befor you are married, even though I don't think you shoud, and the Bible says you shouldnt, but because that was the first example that came to mind. I just that if you see thing's that happen to other people, and the thing effect their life negatively, maybe it will have a negative effect on you so you shoild avoid it. T.V. series often show people hitchhiking and getting picked up and molested, or other things to try to make people realize things to watch out for. If you pay attention to these things to you might have a happier life.

In his discussion of the research session during which we discussed Albright's essay, Robinson indicates that one of the readings of the essay recorded in several of our protocols was that of the "prescriptive grammarians" among us who noted "the applications or mis-applications of absolute rules"; another was that of the "liberal grammarians" among us who, moving from prescription to description, measured "sentences against templates formed from the conventional usages and conventionalized expectations of favored groups—those with prestige and power." The protocols of both the prescriptive grammarians and the liberal grammarians recorded the nonstandard usages in Fred Albright's essay, such as: "the spelling of 'befor', the use of apostrophes in

third-person singular verbs and the failure to use them in expected places (wont,didnt,shouldnt) [and] the spelling of rather than have for the reduced auxiliary verb in line 9." A third reading Robinson notes was that of the "critical thinkers" among us who, when examining the patterns of argument in Albright's essay, found that he provided "grounds for his articulation rather deftly by glossing the statement in such a way as to make more concrete the notion of what it is to learn from experience: to learn to pay attention—to see not only an event but also the consequences of that event in order to adjust behavior accordingly." Robinson also indicates that the "critical thinkers" among us found fault with Albright at the beginning of his third paragraph when he incautiously explains that his example just happened to come to his mind, that he had not intentionally chosen the example he presented. Careful reasoning would not have allowed him to argue his case with just any example that came to mind.

As he reviews the readings we were able to construct of Albright's essay in our research session, Robinson argues for one that emerged, one that could not have surfaced had Albright's text been examined in isolation, as an artifact, by any other group of researchers. Because Albright's classroom teacher was a member of our group, she was able to explain to us why he had included the seemingly "errant move" in his argument: Albright had come to be known in his classroom, where discussion was encouraged, as deeply religious, as "fundamentalist" in his "approach to human conduct" (p.13). This stance was a problem for him with his peers. Reflecting on the student writer's rhetorical situation in his essay, Robinson summarizes a lesson of which we were reminded dramatically by Albright's essay in our research session: Like Albright's, "all effective writing is similarly centered in a writer's anticipation of the needs and interests and expectations of real or imagined individuals whose readings will count for something in the writer's system of values." We saw before our eyes in each other's evaluations of his text how reading Albright's essay in the context within which it was composed required that we value it quite differently from individuals who would not have that opportunity. Informed by his teacher, our research group came to see Albright's argument as well-constructed, along classical lines, within the rhetorical community in which he was writing because his thesis and support (logos) were as sensitive to the values of his audience (pathos) as he knew his readers would be sensitive to his values (ethos).

Although the strategies we employed to study the student papers varied, all of them provoked discussions like the one that

emerged from our readings of Fred Albright's essay. A "read around" strategy worked particularly well for us as we pushed toward an understanding of what we wished to call good writing. To realize this strategy, those of us in the "school writing" group divided ourselves into subgroups of three or four teachers in which we read and evaluated four or five essays from a set of student papers that one teacher had brought for study. After identifying the papers we thought most successfully fulfilled a given assignment, we exchanged all our papers with another group. After each group finished reading all papers, we noted those we identified as most effective and worked together to name as criteria the characteristics and qualities of the writing that we were valuing. We repeated this procedure frequently with different kinds of writing, noting that what we valued changed depending upon the writing we were examining. The "read around" strategy made its way back into most classrooms, where teachers most frequently used it to help students move from rough drafts to final versions of their work. In most cases, after student "read around" groups identified criteria for a "good paper" on a given assignment, students revised their papers in light of the criteria that they and their teachers agreed would be the criteria for evaluation.

Through our collaborative inquiry into the characteristics of the papers we were reading and into the sometimes private values that shaped our readings, we began to mold ourselves into an interpretative community whose values could be made public, made subject to critique and criticism, so that they could serve as our community's shared values. The building of such a community is essential if assessment is to work. It is no less essential if teaching is to be made purposive and coherent, if teaching is to be aimed at rationally defensible goals designed to benefit the whole teaching-learning community, students, teachers, and those who would administer schools well.

The "Other-than-School" Writing Group

Members of the "other-than-school" writing group began their work by involving a number of their classes in letter writing exchanges. Students in some classes exchanged letters with students in other classes in Saginaw; students in other classes exchanged letters with students in classes in other cities and states. Because we decided that students corresponding within the city of Saginaw should be assured anonymity if they wished it, we asked those students to choose pen names for their letter writing. This practice was not adopted when students wrote to pen pals outside their own city.

Assuring students that they did not wish to invade anyone's privacy, teachers explained that they would be studying the letters students exchanged to discover how students' writing took shape when it was not composed to satisfy school assignments. Informed by research into students' letter writing conducted by Shirley Brice Heath and Amanda Branscombe (Heath, Branscombe, 1985), we hoped our students would learn from one another during their correspondence; we also hoped that by reading over their shoulders, we would learn what our students could do when their writing was composed for purposes of real communication with others real to them.

In our study group sessions, we used a teaching strategy adapted from a "Talk-Write" practice (Wixon, Wixon, 1977). We conducted thematic analyses of students' letters to learn about two things: the topics that interested students and the strategies students employed to establish written conversations with each other. Working in pairs, members of our study group would read a set of letters exchanged between two students. One of the pair would tell the other what she or he was noting in the letters; the other would write these comments on newsprint and then rehearse them for the first to make sure she or he had heard correctly. This "Talk-Write" strategy quickly made its way back into most teachers' classrooms as a device for encouraging students to revise for amplitude and clarity.

A brief analysis of an exchange between an eleventh-grader at Saginaw High School, who chose to write under the pseudonym "Darryl and Joe," and a tenth-grader at Arthur Hill High School, who writes under the name "Monique Bailey," illustrates the features of students' letter writing that we noted for one another in our research group. In this exchange Darryl and Joe (hereafter Darryl) was the first writer. Like many Saginaw students, Darryl used his pseudonym as a topic with which to begin conversation:

Dear Monique Bailey,

My name or my pen name is Darryl A. Joe. I got the idea from Run D.M.C.'s song called Darryl and Joe. My teacher said we had to mention our school which is Saginaw High but you already should know that. I went to Arthur Hill last year and I had fun but it nothing compared to the things I do know. I would tell you but the teacher might read this and find out that I'm not the well-behaved, well mannered person she think's I am. Right now you are probably saying "if she read it you just told on yourself" and you probably added some more pronouns. But I'm not.

In a short summary of myself I'm six feet two inches, I have short hair, brown eyes. I wear size ten and a half shoes and I like to

have fun, go to parties, "chill with my boys." And I don't like talking on the phone because it seems as if the women that you're talking to is blowing her breath on your ear and making it sweat.

I also have to write about my feelings about the program of writing in these letters. I like it because now I can ask you about certain fact or lies that people told me about what happen at Arthur Hill, and if thin work out maybe where the "Hill-lites" are hanging out this year, so my boys and I can come check you out.

Next week,
Darryl and Joe

? about U

Answer from me

-wht grade are you in?	I'm in the eleventh
-how old are you?	I'm sixteen
-what color are your eyes?	
-what color is your hair?	
-Do this feel like a test?	In a way

We noted that to begin his correspondence Darryl used as topics: his pseudonym, the letter writing project, and his attendance at Arthur Hill, so far as he knew, the only things he might have in common with his pen pal. We characterized his letter as friendly, inviting response, but cautious. He protected himself against an unfriendly pen pal by communicating himself as strong and worldly, describing himself as big physically, as a risk taker, as a young man who knew some young women, a young man who would like to "check out" Monique Bailey at Arthur Hill. We also noted that Darryl used language to talk about language when he wrote, "and you probably added some more pronouns."

When we first noted students commenting on language forms and language use in these letters, we wondered if they were doing so because the letters were written in English classes. We do not know the answer to that question, but we know that we learned something we had not expected to learn: The students whose letters we studied are conscious not only of the meanings of the spoken and written language in their environment but also of the forms that language takes. We found in the students' interest in and sensitivity to language an opportunity for the development of teaching strategies, since sensitivity to the implications of language choice is part of the competency effective writers bring to the creation of texts.

Monique Bailey's first letter to Darryl crossed his in the mail; therefore, she was not writing a response to his letter, rather she, too, was introducing herself to an unknown pen pal:

Hi!

Well, I'm a tenth grade student at Arthur Hill and as you may have guessed I am a girl. I chose the pseudonym, Monique Bailey, because I like the name Monique and my real initials are M.B. Actually, I go by my middle name so many people don't know that my first initial is M. Why did you choose the pseudonym Darryl N. Joe?

When I heard about this project in Mrs. Oliver's American Lit./comp. class I thought it would be an interesting way to meet new people. Meeting new people is one of the things that I enjoy doing I also like shopping and tripping out with my friends. To some people I appear kind of shy but to those I know I am the farthest from shy. I used to be able to say I like school but now I almost hate it. Well enough about me, tell me some things about yourself!

Talk to you next week,
Monique Bailey

In our inquiry, we noted that Monique begins with a discussion of her pseudonym and that her discussion is an especially generative one. She builds into it an invitation to her pen pal to discover who she really is, an invitation that Darryl accepts in his next letter as he makes use of both the form and the content of Monique's letter in his response to it. In the salutation and first paragraph of his next letter, Darryl responds to the form of Monique's salutation; in his second paragraph, he pursues Monique's allusion to her real identity, and then brings the subject back to his own pseudonym and his efforts to learn how to spell it. Darryl not only opens up the discourse, but as he does so, he documents his observation of language forms:

What's Up?

I'm glad that you started your letter by saying "hi" because our teacher would not let us start ours like that. She said we had to start out by saying "Dear....," and it doesn't seem right to say dear to someone you don't know.

From reading your letter I think I know who you are. I am not sure, but I can ask this girl I know that goes to Arthur Hill. and if I am correct I will tell you my name. But if I'm not, I will continue to go by the name Darryl. In my letter, I probably spelled Darryl two ways. that is because I asked these dudes how to spell Darryl and one said one way and the other said another way.

Another trait I got from your letter is that you write very neat. I "needless to say" don't write neat. From your handwriting I figure

that you are very neat or should I say well groomed. I might be wrong but I don't think so.

The remainder of Darryl's letter reads like his third paragraph. He uses something Monique has written or something he has surmised from the form of her writing to speculate about her and to invite her to tell him more about herself. In her response to his letter, Monique responds to his interest in the formal characteristics of their letters. Her comment is a thoughtful response to the uneasiness Darryl obviously experienced when he addressed her as "Dear" in his first letter, and it is an effort to put him at ease. Following this commentary, Monique expands a topic she has introduced before: She provides Darryl more clues to her identity and asks him to give her some indication of who he is.

Dear Darryl,

Well, this time I was also told to start by saying "Dear...." I guess its okay since we are getting to know each other. So you think you know who I am. I doubt if you're right but who do you think I am?

The letters that Darryl and Monique exchange continue to develop the theme of their identities until they finally exchange names. With each letter, new themes are introduced. Monique initiates the themes of community events and mutual friends; Darryl builds upon the themes in his next letter, again making observations about language.

I know two Catrina Williams, but I don't know how they spell their names. So I might ask both of them if they a girl named Denise, who is in tenth grade and goes to Arthur Hill. Then I will say that she might be a Xino. If they say yes then probably come up to Arthur Hill to find you. My friend and I were thinking of coming up to Arthur Hill to meet our secret "ad ires" or to meet our pen pals with sounds more like a pet name than a name for someone you might care for or have feeling for.

In her response, Monique introduces still other themes; and, once again, Darryl expands upon the themes she introduces in his next letter. Eventually, the topics under discussion in their letters include their families, especially the death of Darryl's stepmother which takes place during the letter writing; Darryl's social life, which includes smoking and drinking; Monique's social life, which revolves around a church youth group; the different courses of study they are pursuing in school; and world events, especially

the United States' bombing of Libya which takes place during the letter writing. It is apparent to the reader over their shoulders that these two young people bring different interests and different experiences to their letter writing; it is equally apparent that they understand how to discuss with one another subjects that they have experienced variously. Neither relinquishes his or her identity, but in obvious efforts to be gracious, they use language thoughtfully.

In his next letter, Darryl comments on the fact that Monique is a successful student. He compliments her, and asks her to tell him more about her studies when he writes: "You must be very smart or have an outstanding talent since you go to C.A.S. What do you go their for?"

As those of us in the "other-than-school writing" study group read set after set of letter exchanges, we observed students using language in a variety of ways to accomplish a variety of purposes. We saw students' intentions take shape in rhetorical strategies that often surprised us but always pleased us for what they told us about our students' competencies as users of the written language to accomplish their own purposes. Some students, like Darryl, wanted to meet their pen pals. Others wanted to share stories about the sports they enjoyed; others wanted to learn about each other's cities; others wanted to discuss colleges, musical groups, and so on. Still others, like Champaign Kane, wanted to encourage reluctant writers to respond to their letters.

Dear #12,

I am very disappointed that I never received a letter from you. I'm a nice person you'll just have to find out. Why didn't you write, were you sick or you just didn't want to write back.

I still want to know about Arthur Hill and what it's like because I want to go there next year. I'm depending on you to tell me how it is, but if I never get a letter from you what will I learn? Think about it #12, you'll see where I'm coming from.

Stay a nice person ok? And write back soon.

Your friend,
Champaign Kane

Number 12 wrote back. So did more than one thousand students who participated in the letter writing project. Persuaded by their letters that most students, even those who are typically unsuccessful writers in school, had been able to write effective and meaningful letters to their peers, we decided to test our hands at creating "in school" writing tasks that engaged student writers the

way the letter writing project had. With this goal in mind, we returned our attention to assessment, and, in June 1986, we re-joined the "school writing" group to work together for one week at the close of school to study papers collected by four of our colleagues who had conducted a writing assessment in their classes in late May.

TWO APPROACHES TO ASSESSMENT

Four teachers who had been working in the "school writing" group were interested in testing for themselves the strengths and weaknesses of the assessment instrument and criteria that the Michigan Educational Assessment Program (MEAP) had used to sample the quality of students' writing in Michigan in 1982. During the last week of school in their junior- and senior-high school English classes, these teachers administered the 1982 MEAP test of persuasive writing that asks students to argue for or against the establishment of a recreational center in their city:

Some high school students have proposed converting an old house into a recreation center where young people might drop in evenings for talk and relaxation. Some local residents oppose the plan on the grounds that the center would depress property values in the neighborhood and attract undesirable types. A public hearing has been called.

Write a brief paper to turn in at the public hearing supporting or opposing the plan. Remember to take only *one* point of view. Organize your arguments carefully and be as convincing as possible. Space is provided below and on the next three pages.

For one week in June, 1986, approximately half of our research group came together to study issues raised by the assessment our colleagues had conducted and to study the student writing produced in the 20 minutes allocated for response to the assessment prompt. Convinced that we needed to bring both personal insights and those of known experts to the evaluation task we were undertaking, we prepared ourselves for our work with three activities. First, we worked to reclaim our own experiences as student writers. We recalled in writing for ourselves and in discussions with one another specific memories: memories of ourselves learning to write; memories of ourselves as students pleased with our writing, as students frustrated with our writing; memories of our successes and failures as writers in school. Second, informed by these recollections, we asked ourselves to write under test conditions to the MEAP prompt we had asked students to address, and

we reflected in writing and in discussion with one another on the experience of having written the assignment. Third, we reviewed what scholars who have written about the assessment of writing have said about writing assessments in use at the time (Cooper, 1981; Cooper and Odell, 1977; White, 1985).

The next day we began to study carefully 13 student papers. In our evaluation of the 13 papers, we did not rate the students' writing according to the MEAP primary- or secondary-trait criteria for evaluation; instead, we read and talked about each essay in our own individual and collective terms. To illustrate our discussions during the three days, I reproduce one of the student essays and two teachers notations of the talk that surrounded it. In our study group, we referred to the student essay as Paper Q:

I present to you: a story of the Fridays and Saturdays spent at the newly reconverted recreational center.

It is a pleasant Friday afternoon. School had just ended, and the first arrivals at the Saginaw teen Recreation Home sit on the porch and on the lawn in the brilliant May sunlight reading or laughing. All in all, a picture of adolescent camaraderie that sparks memories of your own school days.

Then, glancing at watches and gathering up books, the kids start home for dinner. As you watch them leave from across the street, you think that the idea to convert the fire-trap to a teen social-spot might just work. And so, with high hopes for today's youth, you sit at your table in the dining room, and enjoy your dinner.

And so all goes well that night, except for a car that apparently had lost its muffler, coming by about midnight. And the next day is even better. Beautiful, pre-summer weather, the kids are out across the street playing frisbee, you finally mowed the lawn, it's a great day to be alive!

As always, the kids break about five o'clock, to go home for dinner. But unlike other nights, some come back. They mill around the porch area shouting back and forth to each other. Suddenly, the language turns to obscenities. Polack jokes, racial slurs, and slang terms for female anatomy fly through the air like dirty birds. You furrow your brow thinking "how childish," and go inside to watch "Richard Pryor in Concert" on HBO.

Well, Pryor wasn't that funny, so you go to bed early amidst the cries of juvenile perverts. You are awakened along about eleven-thirty by the sound of breaking glass. It comes from across the street.

You rush to your window, and across the street, you see three full carloads of kids pull up, tossing beer bottles from the windows, screaming out, "motherfucker!!" for no apparent reason, and just making a hell of a racket.

You watch dumbfounded, as the children climb from the cars. Leather, denim, and spikes meld together in a terrible sadistic-macho mishmash. You think to yourself, "It's only a matter of time." And it is.

Two teen-age boys square off in the middle of the road, ready to go at it, over something stupid probably. They meet under the street light, and begin beating each other. You think to yourself "Thank God they don't have knives." And, determined to put a stop to this, you open your front door, and bellow, "Hey!, knock it off!! Get out of here!!!"

From the back of the crowd you hear a voice yell, "Hey man! I got my shit on me, you want me to take you down, man?!!"

Two gunshots ring out, and you dive for the phone to get the police.

And as the cop pulls the last kid from the near-riot, he looks down and says, half-angry half-with pity "What are you here for kid?"

And the kid looks up with pure innocence in his eyes and says, "rest and relaxation."

One of the teachers who took note of our discussion of Paper Q recorded these comments:

Wow! I can smell and hear in this paper.

I love it. Richard Pryor in Concert is such a great contrast which is so similar to the language of kids.

It is so descriptive and cinematic. The imagination is wonderful. It is like a parable or a camera.

The audience is invited into the story. It felt like "Twilight Zone."

He never loses his point. He persuades so well with a story. He follows all of the rules. He has developed arguments against violence, not only in the conventional manner. He takes a position. He talks about good and bad teens by giving illustrations. This comes out even by the use of night and day.

We all love it. We all think that it is the number one paper.

It is also basically correct in the surface features.

This student must love to write. He is gifted.

You can see in his writing that he is turned off by convention.

This is narrative writing. Can we score this type of writing by using the same list of criteria?

I think of this kid as an acned Caulfield. He is potentially a saint or a demon.

What difference does it make what kind of kid he is? He writes marvelously. I found him caustic and bitter at times and it seems that he

is disdainful of his peer group. Perhaps he may be a social misfit. But, what difference that makes in how I grade his paper is beyond my comprehension. The kid is good. Whether he is emotionally unbalanced or not, he pulls my chain. God, look how many famous writers were off the wall. We have to look at the concrete things that the paper contains. It is correct, unusual, and above all, we all found it very interesting and spell-binding to read. He goes beyond just being effective, to me.

Another noted these additional comments:

showing not telling
reading into the text: what different people in the text represent
what the writer captures in the dialogue
Is it the vocabulary that impresses us?
audience is invited into the story
the metaphors, the imagery, the irony

At the conclusion of the June meeting, we rank-ordered the 13 student papers we had read and discussed impressionistically together. Paper Q led the list.

During the course of the June meeting we cohered as a working group. We had come a long way toward developing a common sense of the challenges our students faced as writers; we had come a long way toward understanding those of our practices that enabled and disabled our students. And in terms shaped by one of us, we had a generative formulation of an issue that perplexed us and that we wished to pursue in subsequent meetings: As teachers who must evaluate students' writing, should we evaluate the polish or the potential that we see? As teachers who must both encourage student writers and prepare them for others' expectations of their writing, should we evaluate polish or potential? We know what large-scale testers do, but what is it that we should do?

Those of us who attended the June workshop and evaluated 13 student papers impressionistically asked participants in our project who had not been able to attend that meeting if they would convene early for our scheduled August workshop in order to evaluate those same student papers according to the sets of MEAP primary-trait and secondary-trait criteria. We made this request of our colleagues because we wanted to compare two kinds of talk about the papers: the kind that emerges when readers are trained to evaluate papers according to pre-determined criteria and the kind that emerges when readers are left free to find words to name judgments that have been reached intuitively.

In preparation for its work, the August evaluation group met with the Director of Assessment for The University of Michi-

gan's English Composition Board. He trained those who attended to use primary- and secondary-trait criteria for assessment developed by MEAP. Because we had by now reached a level of comfort with and trust for one another, we were able to tape-record our talk from this meeting on. The following discussion about Paper Q demonstrates teachers' readings of the paper. The talk is much like that which we recorded in notes in June, with one exception: We were laboring to make our readings of the paper fit the criteria; or, to put it another way, we were laboring to make the criteria fit our readings.

BU: I wanted to give it a 1 when I first read it.

BV: What did you give it in the end?

BU: 3.

BV: Why did you want to give it a 1?

BU: Because it seemed to me that it didn't have a position.

RB: Should we read Q together before we talk about it?

[Paper Q is read aloud.]

BU: But, does it go back, does it have a clear cut position?

LB: I think it was very clearly against it, and I also think that it addressed—I gave it a 4 because I think it addressed both the opposition and their own opinion. That they were saying that, "Yes, this is ideally what it could be like. Yes, I understand that this could be very positive things for the kids, but in most instances this is what the reality is.

SF: And even though they don't state their position, it's there. They'll state it, I'm saying, in the introduction or in a sentence here. Without a doubt you know the position.

JJ: Yeah, there was a parody on *The Twilight Zone* in the opening paragraph. I present to you.... You could almost hear Rod Serling. (Laughter and chatter). *The Night Before Christmas*. The images are—at least in the positive part at the beginning—are almost parodies of our adult memories of what it was like in childhood because I looked at Sharon, and we both had that strange look. Like it was when we were kids? And the obvious negativity about the reconverted house at the second half, I thought that came through clear...I would have been put off by the words, I guess, in another context, but I was caught, I think, by the parody and the satire in the situation and even was willing to

overlook those, probably because they're consistent with what my framework is about, the group she's talking about.

PS: Don't you think they were used very much for their appropriateness in that context? Intentionally picked in other words they were colloquial slang just dropped, they were....

SF: I think that this writer really has a style for telling a story. It gets you caught up.

JJ: But Bea, I agree with you. I figured there is a percentage of the population to whom that would be ambiguous and would probably be offended by the language at the end in regards to her argument, be upset and waiver between the issue of pro and con. So I gave it a 3 because of the ambiguity that I anticipated in the general population. If it were in class, I probably would have given her an A and warned her very, very heartily about her language.

PS: Would you really, Jim? You would really say something to the writer about using that language in speech? How could this writer have communicated what the writer's trying to communicate without doing that?

JJ: I think there's probably a general term that could have been used instead, I mean she already said....

PS: They cursed? (General reaction.) But, I mean the whole piece is concrete as opposed to abstract—is one way to put it or for the immediate specific example as opposed to the generalization—so that everything here if it's to be consistent and integral and so forth—this is the sort of thing I'm advancing—has to be concrete and therefore to make the point of sunny afternoons when they're sitting there opposed to—night falls and the whole world changes—is to be just as explicit in your language as Chaucer or Shakespeare.

RB: I gave it a 3 for a different reason, not to change the subject. I thought that this is clearly a pretty good writer. That his/her ability to describe a very poignant scene—and I thought there was quite a little bit of slang and so forth—but it seemed to be generally well chosen, so I wasn't inadvertently pulled to it. What I thought was the problem—and I thought most people would give it a 3—was that it does address both sides of the issue, but it didn't seem to me to develop a whole lot. That's to say that it's very poignant in evoking a kind of scene, one which is idyllic and that the center is going to be a very swell thing because kids are basking in the wonderful May sun having lemonade (laughter) and then there is a street fight. So it's clear where the author's position is that they don't want the center and so forth, it was dangerous. But when you get right down to brass tacks, I'm just thinking now sort of the bureaucrat type with the criteria to deal with, it

doesn't say much about the reasons for or against the recreation center. It simply invokes two very different kinds of moods or very different scenes, and I'm not sure that, at least the question that went through my mind was whether you want to call that development.

CW: But I think that person did that for a purpose. If you go to the first couple of paragraphs, that is the idealistic viewpoint there, but at the end you come to this negative viewpoint, and the person is really saying to me—the way I interpreted it—whether it is good as it is in the first two or if it is bad as some of the last paragraphs, the negative aspect is going to outweigh the good. And what he's really doing here is saying, Yes, it could be this way, it's good, but any time any one little thing negative comes up, we're going to play that up much higher than we would the good part. And therefore, it should be this way, the first two paragraphs, it should be this way. But if anything happens, it's blown.

PS: Can I argue for development too, and hopping on what Carol said? We are presented with the idyllic scene and the language, like—it's Saginaw Teen Recreation Home. It's a home, not even a recreation center, it's a home. And brilliant May sunlight, reading and laughing, adolescent camaraderies. Then, glancing at watches so that "then"—the new paragraph—the turn. We get the first hint of another mood, but it's only subtle, the muffler. Back to the next day. *Then*, so as I see one line tapering, the other line is building, but they don't happen in absolute contrast to one another. At the same time, if you have the idyllic May sunshine, you have this little rumbling in here. And at the same time, that that takes over the other thing is still there tapering, and I think that quite subtle development. For 15 to 20 minutes for a high schooler.

RB: But it's a manipulation of emotion, I'm not sure if it's development.

RH: As a reader of argument, when we use narrative, you expect for them to step aside from the narrative and elaborate more on the significance of the narrative and this does that but it just does it in one liners and it does it in very symmetrical fashion, I think. A picture of adolescent camaraderie. He's telling you what he's presenting in other words—or she—and then down here it says "a great day to be alive." There you have two positive comments on the scene. Then he comments it's only a matter of time, and it is.

BV: As Jim says, all controlled cliches.

RH: Very controlled cliches. And then you come to the end. Rest and Relaxation. I mean the use of that just minimal commentary stepping outside of the narrative is so effective, but it's through that commentary that you're looking for to score under the primary trait guide and it's

something that hangs in the air rather than on the page for me as a reader.

BV: Well, then why wouldn't it be a 4?

LB: It is for me.

BV: I mean if we take a different point of view that he or she is making this statement. But it would be a 4, wouldn't it?

JJ: Back to Bea's point. I think this is a more artistic treatment of what was supposed to be an expository writing assignment and the artistic treatment, that compilation of images and some concepts in there and so on, is confusing. You are not sure exactly what the point is, whether the first couple of pages and paragraphs are supposed to be what it's like, as Carol says, and then she destroys what it might go to poorly, clumsily, by throwing in the swear words and all the rest. You're just not sure with the highly artistic treatment I think this is, of a vital issue, community issue. Not you, but people generally or a group of people. I consider this more art than I do exposition. I consider it more on the descriptive side than even on the narrative side. It doesn't tell a story so much as it describes images and so on that she's using to try to convey ultimately, I think that the reconverted recreation centers or home for young people are probably a poor idea.

Even as it illustrates generally the texture of teachers' talk about students' writing that has characterized all of our research sessions in Saginaw (the emergence of the different values and expectations teachers in one community bring to bear upon the task of evaluating their students' writing), this transcribed discussion illustrates more particularly how rater-evaluators work to fit their readings of a particular paper to predetermined criteria. Is Paper Q persuasive? Can it be called effectively persuasive if it is narrative and not expository? Can it be called effectively persuasive, can it be given a 4, if it does not "systematically define and defend a point of view," if it does not "present at least two moderately developed lines of argument, one which supports the position and one which answers the possible arguments raised by the opposition"? Can arguments and positions be implied or must they be made explicit in order to "systematically define and defend a point of view"?

These are the kinds of questions that talk about papers focused upon preset criteria inevitably provokes. Readers obliged to employ them inevitably struggle toward a way to make their intuitive judgments consistent with features named in the criteria. It is difficult to say that better judgments emerge from these

struggles; it is interesting to note that Paper Q, which was rank-ordered first among the 13 papers by the June discussion group, was rank-ordered ninth among the 13 by the August discussion group. Something new, and perhaps something alien, had been introduced into the interpretative community we were forming in Saginaw.

Although the August discussion group appreciated Paper Q, and saw it as an imaginative and effective response to the task of persuading an audience toward taking action, the majority of us remained convinced that we could not award it a high rating according to the MEAP criteria because the essay violated the generic expectations embedded in the descriptors used to express those criteria. Our intuitive judgments that Paper Q was an unusually outstanding text could not find expression in the discourse imposed through the descriptive language of the MEAP criteria. In developing our own criteria, our own method for allowing students to show us their competence with the written language, we decided we did not want to limit our ability to recognize and reward unusual or unconventional solutions to the problems we would be posing for students.

Reunited with our colleagues who had attended the June workshop, we reformed ourselves as a group with a mission: to plan a writing assessment for tenth-grade students in Saginaw. Encouraged by their summer research, every one of the 20 school teachers in our group began to conduct inquiries in his or her classroom in order to learn from students what kind of writing assessment would best enable them to compose as well as they were able. A description of just one of these inquiries illustrates the contribution students made to the assessment we finally conducted.

Students Join The Research Project

In the fall of 1986, one teacher invited her eleventh- and twelfth-grade students to join her in a four-day research project. She explained that for half of their class hour on each of the first three days, she would present a writing assessment prompt and its attendant criteria for evaluation; for the second half of the class hour, students would try their hands at writing a successful essay to satisfy the prompt and criteria. On the first day, the teacher presented students a prompt asking them to write about unfairness and attendant criteria developed in 1985 by the Ann Arbor Teachers' Assessment Committee; on the second day, she presented a prompt developed by The University of Michigan's ECB to place students in the University's writing program and its attendant criteria (the

prompt, which asks students to write a letter to their local newspaper indicating how serious they think the problem of smoking is and how it may best be solved, requires students to begin their essay with two given sentences); on the third day, she presented students the MEAP prompt on the community center. On the fourth day, she asked students to write to her telling her which writing tasks allowed them to compose their best writing and which did not.

When she brought her students' essays and critiques to the next meeting of our study group, participants who had experience with all three prompts and criteria read the students' essays and their commentaries. This study of the students' papers led us to conclude that the quality of writing across the set of a student's papers was different as often as it was consistent. Informed by the teacher who conducted the inquiry, we learned that the student writers who composed sets of well-written essays were not necessarily students who were writing class assignments successfully. We also noted two other things that interested us: Almost every set of papers had at least one effective piece of writing in it, and students' written reflections about the inquiry in which they participated were consistently thoughtful and thought provoking. These excerpts from two students' reflections illustrate why we found them interesting:

Student Reflection #1

I have this to say this is a special moment for me. Never before have I been *asked* to tell someone what I thought of someone's assignment. Oh, boy, this is going to be fun, fun, fun!

Well, where do I start? The criteria used to judge a good essay. I have to say that the criteria pretty much went in one ear and out the other, because, in the course of my essay, I manage to ignore them all anyway. Why? Well, I figure that the people reading all these essays will get tired of reading the same thing over and over, so I try to take a new angle on the subject. In doing that, I've found I often break the rules set down for me. Oh, well.

The subjects. In a free style writing evaluation, I can write anything. Anything at all. What I *didn't* like was The University of Michigan's prompt. If you're going to give someone the first two sentences of their essay, why not give them the last two? Then why not give them everything in between, too? Just a pet peeve, I guess.

Well, that's all I have to say. I hope I've helped you out. If not, thanks for listening, and thanks for letting me spill my guts, as it were.

Student Reflection #2

First, I don't understand what is the purpose of having students respond to these prompts. Are the students being judged on their ability to write something cohesive and argumentative in a short period of time; *or* are the students being judged on their creativity? Now, I imagine you want me to define creativity. Let me begin by saying what it is not. Creativity cannot be found when given pages of criteria before writing anything. I find that when I know exactly what is expected I temper my thoughts to fit the criteria and work on grammar and mechanics, thus the writing becomes inconsequential. I think everything I have ever written is inconsequential and probably everything I will ever write. But I am digressing (something you are not supposed to do according to The University of Michigan). Creativity is the act of creating, expressing something from inside. Every topic we wrote to appeared trite, thus everything I wrote had already been said and lacked importance. What would be a prompt to better test creativity? A prompt without a question or an argument. A statement and then the word, respond.

If, however, the purpose of these essays is to test our ability to write something cohesive and argumentative, essentially to write something practical, The University of Michigan prompt about smoking was the best. It was clear and provide for uniform essays from all the participants which allows more people who judge the essay to compare them more completely.

The Ann Arbor prompt was simply to big. Everything in life is unfair, but that isn't a good thesis statement for a twenty-minute essay.

The MEAP prompt was, and I hate to say this, boring. It was entirely too predictable and would have been easy to write and say exactly what the readers want to hear.

The most effective prompt for purpose of judging practical writing ability in a large group of students is The University of Michigan prompt. However, I think all prompts failed to effectively tap both the writing abilities and creativity of students.

During a visit two other teachers made to the classes involved in this inquiry to talk with the teacher and her students about their experiences writing the assessments, students indicated that knowing and discussing the criteria for evaluation before they wrote was not helpful to them; in fact, for many it was disconcerting, even upsetting. One young woman suggested that when the actual assessment took place students might be given the criteria on which their writing would be evaluated after they had written drafts but before they composed the final versions of their essays.

We incorporated this woman's suggestion in our assessment procedure.

In another of our investigations into what kind of writing assessment would best enable students to invest their imagination and effort, one teacher in our group asked her students to generate topics for writing that would interest them. The students suggested 37 topics. After narrowing the list down to six, When to Say No; Racial Discrimination; Teenage Stress; Teenage Drinking; Preparing for Interviews; Pressure of School, the teacher's students decided that "Teenage Stress" was the topic they could write about most effectively. Those who voted for it argued that it encompassed the others, explaining to their classmates that if someone wanted to write on preparing for an interview or teenage drinking, she or he could do so within the topic of stress. Having asked her students to test their topic by writing an at-home essay on it, the teacher brought the essays to our next meeting. Recalling the power of our students' letter writing when they were in control of the topics about which they wrote, we were pleased that students had identified a topic for the writing assessment. But as we read students' essays written in response to the topic we had mixed feelings. The writing was competent, but lifeless. In almost every case, students' writing was general and abstract when the topic invited concreteness and specificity. This paper written by one of the students illustrates the kind of writing they composed in response to the topic they suggested.

When you first enter high school, everyone affiliated with orientation encourages you to become active. There are clubs, teams, and services that are purposely inclined to enrich your high school experience. It all sounds tempting, so you go for it.

So now you're involved. It is surprising to realize that you have acquired so many acquaintances. After a while, you have become a very popular person. Everyone knows you from one thing to another. This prestige sort of makes you feel good.

However, gradually your entire life changes. You become "Mr./Mrs. Do-body." There are meetings, practices, or rehearsals, programs *plus* homework and housework! You ask, "Where does social life come in? It doesn't. Your activities are your social life. There is no time for anything else, because you've made a commitment to each activity. If you put your all into everything, then others become impressed and place "honors" upon you—meaning that your name appears in programs that you know nothing about until they day before.

This is where stress comes in. After a while you mentally and physically cannot meet your demands. Somehow it seems impossible! On occasions your activities may conflict. Then what? Something is neglected, obviously. This can result to the feeling of discouragement.

A lot of times you may want to give up. This thought strongly sticks in your mind, which makes situations successful, (sic) take life one day at a time and chaos will calm down.

Although we decided to adopt the topic students had identified for us, their essays provoked us to develop and test with other students several prewriting activities we hoped would encourage them to write with both increased specificity and deeper reflection. Having done so, we gathered for a daylong workshop in which we not only studied the writing several classes of students had composed in response to a variety of invitations to write about teenage stress, but also tried our own hands at writing to the activities we had developed. This led us to give final form to four days of activities that constituted the writing assessment we had worked with our students to develop.

In several subsequent meetings, referring continually to writing students had composed for us as well as to the values and expectations for students' writing that we had identified in our conversations during our study, we outlined the criteria we would use to evaluate the writing students would compose on the fourth of the four-day writing assessment.

THE ASSESSMENT

On the first of the four days given to the assessment, which was conducted in April 1987, in their own classrooms, tenth-grade English teachers in Saginaw prompted their students to write an account of some incident in which they or someone they knew had experienced stress. Following this writing, teachers asked students to share their stories with two or three of their classmates. As they listened to each other's stories, students were asked to note in their test booklets other stressful incidents of which their classmates stories reminded them.

On the second day, students were asked to compose an account of another stressful incident, something brought to mind by accounts told the previous day by other students. In addition, "language lists"—lists of words and phrases that seemed best to capture the notion of stress—were collected on chalk boards in their classrooms.

On the third day, students were given the following prompt and asked to take about half an hour to write a response to it:

Writing Assessment Prompt

Teenage Stress

"Get off my case!"

"Get outta my face!"

"Just forget it!"

At some time or other, most teenagers have observed or experienced stress. Have you ever felt life was coming down on you pretty hard and you didn't know how much more you could take? Think about a time when you or someone you know felt stressed, frustrated, or uptight.

Write an essay based in your own experience that tells Saginaw English teachers what stresses teenagers feel and why teenage stress is a problem.

You may begin your essay however you choose, but if you need suggestions for how to begin, here are several:

Sometimes, I lose my cool.

There are days when I just can't stand it anymore.

The adults I know seem to think that a teenager's life is free of pressures.

When students had written for half an hour, teachers asked them to set their essays aside and turn their attention to the following letter. Teachers explained that the letter would tell them how they might help a partner to earn the best evaluation possible for his or her essay. The letter read:

Dear Student Writers:

As you and your partner read the first drafts of each other's essays for the purposes of helping each other write even better final essays, it may help you to know that your final essays will be evaluated by two English teachers who will each award your essay a numerical score from 1-4.

4=Excellent
3=Proficient

2=Interesting but Flawed
1=Unsuccessful

As you underline the parts of your partner's essay you like and ask your partner questions about aspects of his/her essay, keep in mind these questions that teachers will be asking themselves as they evaluate your essays:

1. Has this writer made clear to me his/her own understanding of what stress is?
2. Has this writer illustrated his/her own understanding with specific, appropriate examples?
3. Does this writer illustrate his/her ideas and illustrations in a way that allows me to understand them as he/she does? That is, are the ideas and illustrations organized so that I can follow them and understand the connections between them? Are they written in language that is clear and interesting? Are there so many usage and mechanical errors in this essay that I cannot understand what the writer wants me to understand?

With these questions in mind that teachers will be asking themselves, offer your partner the best help you can as you underline and question what he/she has written?

GOOD LUCK TO YOU AND YOUR PARTNER.

Students then exchanged papers, underlined sections of their partners' papers they particularly liked, and wrote questions to their partners asking about what was unclear or understated in their papers.

On the fourth day, students were asked to revise or write another version of their essays. It was this final writing that we teachers rated. In two-day evaluation sessions, we read each student's essay a minimum of four times and, in the case of a number of essays we used to monitor the reliability of our readings, a maximum of 20 times.

Below are descriptions of the rankings we assigned students' writing:

Criteria for Evaluation

(4) EXCELLENT (The essay fully engages the reader because of its originality of thought and its accuracy, appropriateness, and freshness of expression. Overall unity and coherence are evident. The writer assumes a clear and definable stance toward his or her materials, makes a claim, and develops the claim through appropriate illustrations or particular demonstration. There is shape and clarity in the organization; functional units (introductions, illustrations, elaborations, etc.) are marked or obvious. There is appropriate variety both in sentence

structure and in the vocabulary employed in sentences. Technical errors either do not occur, or are so rare that they do not interfere with the writer's message or the reader's engagement.)

(3.5) EXCELLENT TO PROFICIENT

(3) PROFICIENT (Although the essay engages the reader, it may lack originality of thought or it may demonstrate weaknesses in expression, development, coherence, clarity, or correctness. Clarity of claims, soundness of thought and development, clear organization, and general correctness of expression are all required for an essay to be ranked 3.)

(2.5) PROFICIENT TO INTERESTING BUT FLAWED

(2) INTERESTING BUT FLAWED (The essay engages the reader, although not fully, because of failures in imagination, inconsistencies, inaccuracies or inadequacies in development, imprecision or staleness of expression, and /or serious problems with the requirements of correctness. Illustrative problems might be:

Thought and Development: Superficiality of thought; reliance on facile generalizations; incorrect statements of fact from which inferences are drawn; irrelevant statements or evidence; needless repetition; inaccurate use of expressions and words.

Organization and Style: Lack of discernible organization; lack of coherence; absence of transitional sentences, phrases, or words. Inappropriate and uncontrolled sentence structure; inappropriate level of usage and diction; vague reference; dangling elements.

Technical Errors: Sentence run-ons, misspellings, agreement errors, missing or inaccurate paragraphing. Errors like these must leap out at the reader and cause the writer's message to be lost for an essay to be rated 2.)

(1.5) INTERESTING BUT FLAWED TO UNSUCCESSFUL

(1) UNSUCCESSFUL (The essay (1) interests the reader in spite of the fact that the reader has to work throughout the essay to make meaning or perceive form or (2) does not interest the reader because of a combination of problems: superficiality or staleness of thought; lack of clarity in the claims; lack of development; incoherence in organization; inaccuracy and/or uncertainty in expression; serious lack of control of the mechanics of written English.)

THE RESULTS

We collected statistical summaries of our evaluations of the writings students composed for the Saginaw Teachers' Assessment

of Writing, but because we were teachers who had taken on the task of assessment, these were not the results we cared much about or made much use of. Results important to us as teachers came from our attempts to find out from students the topics they would like to write about for an assessment and the forms in which they might comfortably shape meanings important to them. Results important to us came as we talked together, argued together, about standards and expectations that we might fairly and equitably apply to the writings of tenth-grade students in a school district like Saginaw's, in a city like Saginaw. What rights do students have to their own language? What rights do students have to a literacy that has to some extent, in their lives outside school, already enabled them in those systems of discourse in which they are active participants? Because we teachers were to be the readers and raters of the writings students composed in response to prompts and in situations that we had designed, talk about standards and expectations had to become very concrete, very much connected to those concrete everyday situations in which students do or do not learn to write and in which we do or do not teach them how.

Still other results issued when the teacher assessors were treated as professionals responsible for reporting the results of the assessment to the Board of Education and to the community. This altered the teachers' roles and their place in the educational hierarchy (they became speakers and not merely listeners in talk among local policy makers about educational achievement). It also gave them a forum in which to raise professionally responsible questions about the enterprise of testing writing.

And then there were these other results as well, the most important of all: Assessors who are also teachers of the students being tested are not and cannot be carpetbaggers. They have to stay at home to face the implications of what they have done, to ask, if only themselves, why some students were able to write effectively for the assessment and others were not. Since the assessment, we have studied the texts students produced in response to it, with special attention to those texts we identified as failures according to criteria we had devised—those texts that did not meet our expectations.

The majority of students who responded to our assessment's prompts composed texts that met our expectations sufficiently for us to assign them the rankings 2 (Interesting but Flawed) or 3 (Proficient). Another number of students, a customary one, impressed us enough by their control of their worlds and their words for us to assign them the highest ranking 4 (Excellent).

The remaining students, also a customary number, were not so ranked; 1 (Unsuccessful) was their score. Given our hopes that we could design an assessment that would enable all students to compose as well as they could, as assessors we studied the texts we had failed: As teachers, we continue to study these texts to better understand why some students failed and where we might have failed them. In our study we have examined all four texts each student composed, not just the text that was scored.

Below is one set of such texts. It is representative, we feel, at least in some of its characteristics, of the challenges failed texts offered us as teachers who would understand how we might have done better.

First Day:

Stress is when you impell force like for example it was this certain person who hang out with this group of people who use to fight and get high and sale drugs And He was allways called a square or a nerd so he was allways thinking should he do it so he could be one of the boys in the crew. So he decided to do what ever they did. So one day they were throwing down at a party so they told him to start it off. So they smoked a joint and he smoked one to but he never smoked before and he started feeling funny. So he walked up to the boy and hit him then they started fighting so then he was given a gun and he shot the boy and then the went to Jail then he could handle it so he killed his self.

Second Day:

Stress

Is when for about two week's I had a lot of things on my mind. It seem like every thing was coming down on me at once. So I went home and tried to rest for a little. But the thought's kept coming back So I got my jacket So it seem like I just exploded and I broke his arm well I was in the police van. I was just thinking about my intire week. And the policeman said I looked like I had a lot of pressure on me. So when I got home I thought I was going to die but after I got all the problem off my chest I felt better. So the next day I seen my girl talking to the boy I got unto it with so then he started kissing So I just went over and they walk away. So when I got to school I could not think in class. So I left school and all most got into a crash.

Third Day:

Stress

One day I had called up a few of my cousin's to go to a dance. And when we were on our way to the party we had seen a lot of boy's they

were looking at us like they wanted to fight so we first kept on riding so we went on to the dance. So as we got inside we were walking around then we seen a few girls then we started dancing So when we walked off the floor the boy's we saw earlier were following us so we went to the bathroom and they seene us and we started to fight so the police officer seen the whole thing so he put them out so we kept on dancing and then the dance was over. When we got out side they had bats and chains, knives. So we all sstept back and on of my counsins ran at the boy with the bat and got his head bust. So we took him to the hospital. Then the next day I felt hurt cause I felt like I should have help him as I could have got hit then I could not even work I felt like life was closing around me.

Fourth Day:

Stress

I remember when we were at the mall and there were 6 boys in the sports store but when we went by the store they ran out and we started fighting and one boy got killed. and that hole day I was thinking about it and I could not work I was in my own world for a long time and we had got locked up for allmost a year and when I got out I could not deal with life.

As assessors, we failed Charles Baldwin: Almost unanimously we gave his texts a 1, seeing them as Unsuccessful. We did so partly, no doubt, for the customary reasons: Given the time he was provided to compose them, his texts were very short; he lacked consistent control of the conventions of standardized written English (although he is a remarkably good speller); he refused (or could not take up) our invitation, as prompters, to move from story to essay, from narrative to reflection, at least not in an assessably successful form. And although we rated only his fourth piece of writing during our assessment evaluation session, in our subsequent research sessions, when we studied all four pieces of his writing, Charles Baldwin's narratives, taken as a whole, posed problems for us too: Most of us read them as incoherent.

It did not matter that as other kinds of readers than assessors, we valued Charles Baldwin's texts as powerful statements about the world in which he lives; we failed them. We did so surely because we feared for his future, in school and beyond, insofar as his future depends upon his own use of language and his tutored capacity to use language. We identified him clearly as a student who is "at risk."

And yet Charles Baldwin wrote, when others certainly hadn't. And later, as we reread his texts as teachers who wished to

help him, we asked ourselves how we might do so. In the stories he composed for us, in the forms he has given his stories, we found openings to spaces within which we might meet him and talk with him, not only about the words he has used and the worlds they have made, but about other words as well: the words he only implies, those he seems to understand as calls for reflexive reactions, and about words we might propose to him, words that can invite him to reflective actions.

Having read the writing Charles Baldwin composed during the third day of the assessment examination, why could we not invite him to rehearse his cousins' talk on the way to the dance, the taunts of the "boys" who wanted a fight, the words of the girls with whom he danced, the lyrics of the songs to which they danced, the commands of the policeman who put the "boys" out of the school, the "boys" challenges to fight, the emergency room staffs' questions, the cousins' answers, the voice of his conscience, the voice of his employer? Having read the writing he composed the first day, why not ask him to speculate about words as causes of actions? Why not invite him to dramatize one of the times when "this certain person" was called a "square or a nerd," to recollect who did the calling? Who looked on? What happened next? Why? Why not ask him to revise the story he composed on the third day by imagining what might have happened if, when he and his cousins were confronted with "bats and chains, knives," they had turned from the fight, figuring their chances for being hurt were too great?

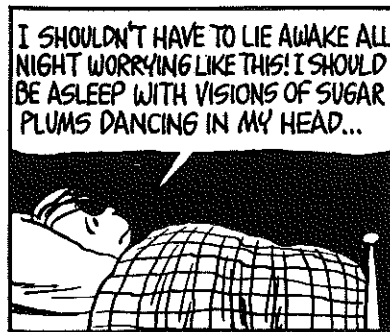
Having read and learned what we could from the texts that Charles Baldwin and other students wrote, those of us who had worked together for two years to assess and interpret the writing competencies of high school students in Saginaw turned our attention to developing a pedagogy for our students' critical literacy. Together, during the school year 1987-88, we designed and taught a yearlong English course that we called Inquiry and Expression, a name we borrowed from TheodoreSizer (p. 132). In our teacher-research project to design a pedagogy that would foster a critical literacy in our students, as in the project we undertook to develop an assessment of our students' writing competencies, we have asked our students to join us as coresearchers.

And in this project to develop a pedagogy that fosters critical literacy in our students, the results of The Saginaw Teachers' Assessment of Writing are to be found. Unlike the charted displays of test scores and the narratives that suggest their meaning that large-scale testers customarily offer as the results of their work, the results of our local, teacher-developed assessment of students' writing are a series of beginnings, the development of

new curricula, the enactment of new teaching methods, the conception of new research projects to be undertaken with students. Those enterprises constitute another story, another chapter in the book that is being written by teacher-researchers.

PART TWO

Assessment Theory



Introduction

George E. Hein

Data banks containing tens of thousands of short-answer science test items are available. In addition, numerous alternative science testing models are now beginning to attract attention: In the spring of 1989, for example, the New York State Education Department required all fourth-grade students to take a science test that included performance measures. But developing appropriate assessment requires more than selecting among the available assessment methods: It requires a rationale. The theory that governs assessment may have a profound effect on the validity and adequacy of the assessment.

The three papers that follow all address aspects of assessment theory as it relates to assessment of hands-on science programs. Although each paper discusses a different aspect of assessment theory, there are common features among them. The three papers affirm the difficulty of justifying any assessment methodology unless theoretical issues are addressed. They further illustrate how primitive our current theoretical base is, how much needs to be done in order to justify any assessment strategy!

Jerry Pine examines the issue through the lens of validity. His concern is not with the various technical uses of the term, but with a larger question: What evidence do we have that any form of testing, whether short-answer, multiple-choice, or performance, correlates with an independent measure of science competence, knowledge, or ability? It is easy to criticize multiple-choice tests as low on face validity since it is difficult to imagine how this form of testing could elicit the information needed to assess students' true science competence. Several authors in this volume provide the reasons for this. But even if we consider performance measures, what independent evidence do we have that high scores on these represent true science competence?

In discussing this issue, Pine points out a further concern: We are missing not only empirical evidence for the validity of our assessments but also agreement on what external variables to use to assess true science competence.

Frank Davis considers the issues from a more general perspective. Different pedagogic points of view exist, based on differing psychological theories of learning (behavioral, developmental, contextual), and these are reflected in science curricula.

If pedagogy can be based on different philosophical positions, and there is a close link between pedagogy and assessment, then we need to examine our assessment methods for a match between the two. Each pedagogical approach requires a particular assessment approach; what we value as the outcome of learning should dictate what we assess. His discussion clarifies some of the issues and illustrates the complexity of developing appropriate assessment models.

Many educators have recognized that current assessment methods emphasize memorization and recall. In an effort to broaden what is assessed, educators have embraced the notion of assessing a hierarchy of intellectual skills. A number of tests of higher-order thinking skills have been devised, and most commercial and state-mandated tests now argue that their products assess these components of intellectual activity. Audrey Champagne raises the sobering possibility that we are a long way from adequately defining these components of the intellect and that many of the claims made for current tests may be exaggerated.

All three papers stress the need for clarifying theory, the close connection between assessment and pedagogy, and the need to articulate how a particular assessment matches pedagogic beliefs.

Finding assessment methods that appear to match pedagogy—for example, performance tasks, portfolios, and other authentic assessments (Wiggins, 1989) in the case of hands-on science that is based on constructivist theory—is only a necessary, not a sufficient, step in the development of a justifiable assessment system. We still need independent data to confirm that these assessments are related to ‘true’ science competence, but in order to collect this data, we first need to decide what that competence consists of.

Assessment and Teaching of Thinking Skills

Audrey B. Champagne

INTRODUCTION

Scores on the National Assessment of Educational Progress (NAEP) Science Assessment have increased slightly since the assessment was first administered in 1969 (Educational Testing Service, 1989). However, the improvement is in young peoples' knowledge about science, not in their ability to apply or to interpret that knowledge. Despite expressions of national concern, the improvement of science-related thinking skills has proved difficult to accomplish. Explanations abound for the notable lack of progress on this critical educational priority. The most usually cited point is the mitigating effects of national testing practices on science classroom practice: Because the science curriculum is driven by tests that primarily measure students' store of information about science, students have little opportunity to develop scientific thinking skills. Underemphasis on thinking skills in assessment instruments is generally attributed to the economics of testing: Exercises appropriate for testing thinking skills are too expensive to administer and to score.

The central thesis of this paper is different. It is that the basic problem is theoretical. Assessment and teaching of scientific thinking skills, as currently practiced, are based on inadequate conceptions of the mental processes that produce skilled academic performance. Conceptions of the relationships between mental processes and academic performance are tacit, idiosyncratic, and untested. The tacit and idiosyncratic nature of these conceptions prevents teachers from examining the conceptual basis of their teaching practices and makes the interpretation of test performance inexact. Progress in improving scientific thinking skills requires subjecting the conceptions that now drive practice to empirical tests. These tests will yield information about which conceptions best explain and predict academic performance and, therefore, provide the theoretical basis for practice.

The development of models that will improve learning requires an intensive research effort. Some preliminary steps in such a research effort and theoretical issues related to them are discussed below.

Before coming to the American Association for the Advancement of Science to direct the AAAS Forum for School Science and the AAAS Project of Liberal Education and the Sciences, Audrey B. Champagne was Senior Scientist at the University of Pittsburgh Learning Research and Development Center. While there, she conducted research on factors that make learning the physical sciences difficult.

MENTAL MODELS AND ACADEMIC TASK PERFORMANCE

Conceptions of the mental processes that determine academic performance are based on observations of students as they perform academic tasks and assessment exercises. Academic tasks—activities teachers assign to students in order to produce learning—and assessment exercises—activities used to measure student achievement—have much in common. Both require students to read and interpret text; solve problems; plan, execute, and interpret experiments; compute; and write. The same type of task—problem solving, for instance—is used to teach and to test. Students both learn how to solve problems and demonstrate that they have learned by successfully solving problems.

Different conceptions of the knowledge and mental processes required for academic tasks lead to differences in instructional and testing practices. A teacher's strategy for teaching complex tasks is based on the teacher's conception of the knowledge and mental processes required to perform the task. When teachers' conceptions differ, the content they teach and the strategies they employ will differ accordingly. Similarly, a test designer's choice of a task to assess a certain thinking skill will be determined by the designer's conception of the thinking skills required to perform the task successfully. If another designers' conception of the mental processes required to do the task are different, disagreements will arise about whether or not the task is a valid test of the thinking skill in question.

Empirically verified models would provide a scientific basis for teaching and testing. Academic tasks could be better matched with mental processes to be learned and unsatisfactory performance could be analyzed to determine what deficiencies in mental processes are preventing satisfactory performance. Furthermore, science achievement could be estimated with greater accuracy if correspondence between item performance and thinking skills was empirically verified.

TAXONOMY CONSTRUCTION AND THEORY DEVELOPMENT

A first step in developing a theory to inform practice is to discover relationships among the many mental processes that educators and psychologists associate with scientific thinking. The strategy employed in this study was the development of taxonomies of science-related thinking skills.

Taxonomy development is a precursor to theory development in the natural sciences. In the descriptive stages of a science, objects of interest are collected and organized in different ways. For instance, early biologists collected and categorized living organisms and early chemists identified a class of substances they called elements which they ordered in a variety of ways. Patterns that emerged as a result of certain systems of categorization and organization contributed to theory development. Biological taxonomies based on structure illuminated patterns and relationships that contributed to the development of evolutionary theory. Theory relating chemical properties and atomic structure was advanced by the periodicity of chemical properties revealed by Mendeleyev's organization of the elements.

Contemporary education and social science are in a developmental stage similar to that of 18th century biology. The professional literature of education and psychology contains a bewildering array of terms for science-related thinking skills. Often these terms are used without definition. Only in reports of empirical research are thinking skills defined operationally, that is, in terms of a task that is used to measure the skill being studied. Nowhere are the relationships among the skills discussed in any systematic way.

The potential contribution of taxonomy development to advancing psychological theory is illustrated using science-related thinking skills that appear in contemporary educational and psychological literature. Terms for thinking skills were collected, then sorted into groups of related terms. The groups were then arranged in different ways to discern what relationships might be posited among them. Three of the taxonomies produced by this process are discussed below (see figures 1-3).

Taxonomy A was the first generated. This taxonomy was generated to ascertain whether a hierarchical organization of thinking skills based on "order" was possible. Taxonomies B and C were generated after conversations with colleagues about Taxonomy A. Taxonomies B and C reflect my colleagues' suggestions for alternative ways of sorting and arranging the terms. No one of the taxonomies represents the domain of thinking skills any better than the others. They are simply representations that illustrate some interesting relationships among thinking skills, relationships that suggest ways in which the domain might be structured to improve theoretical understanding and educational practices.

Figure 1. Taxonomy A

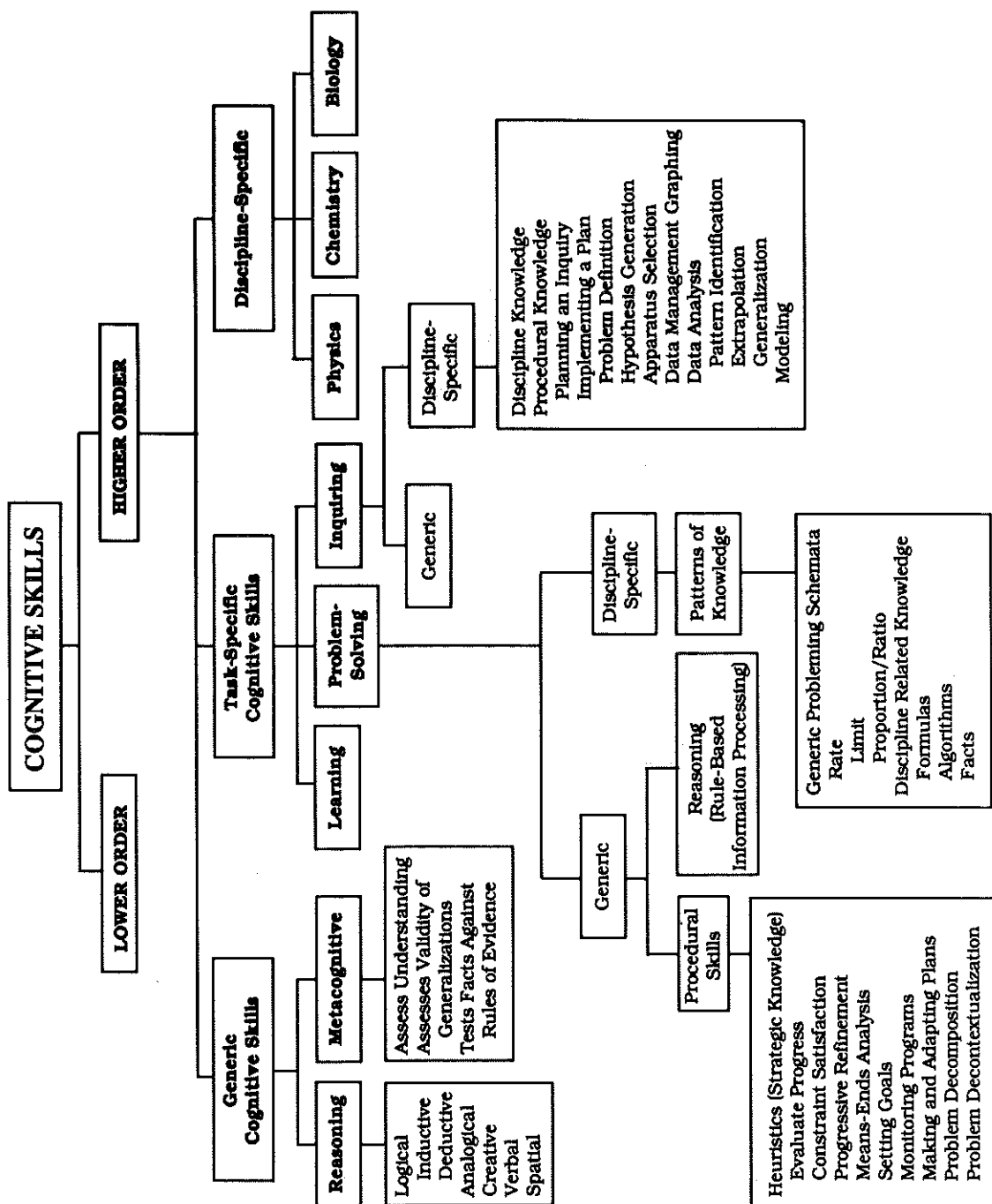


Figure2. Taxonomy B

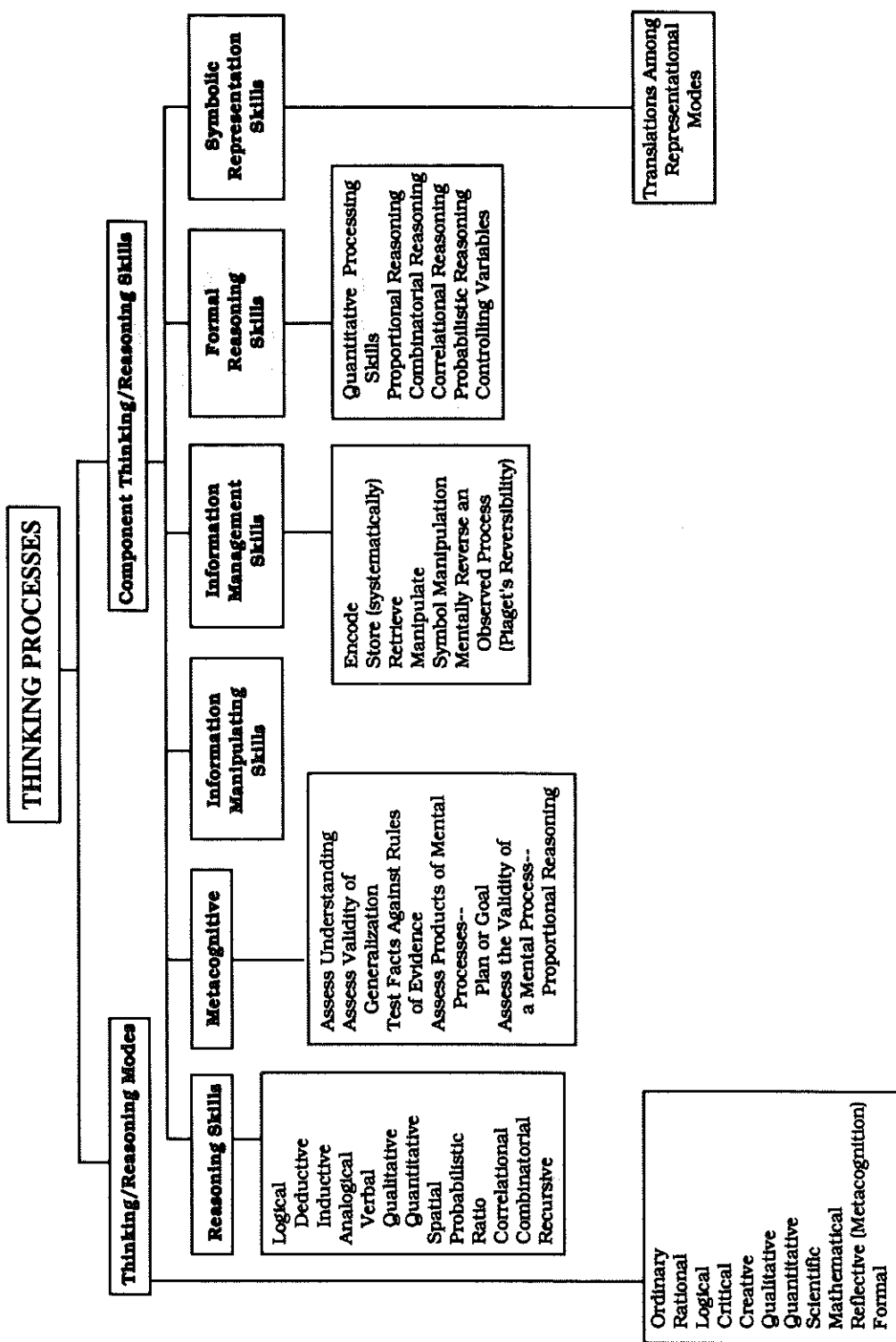
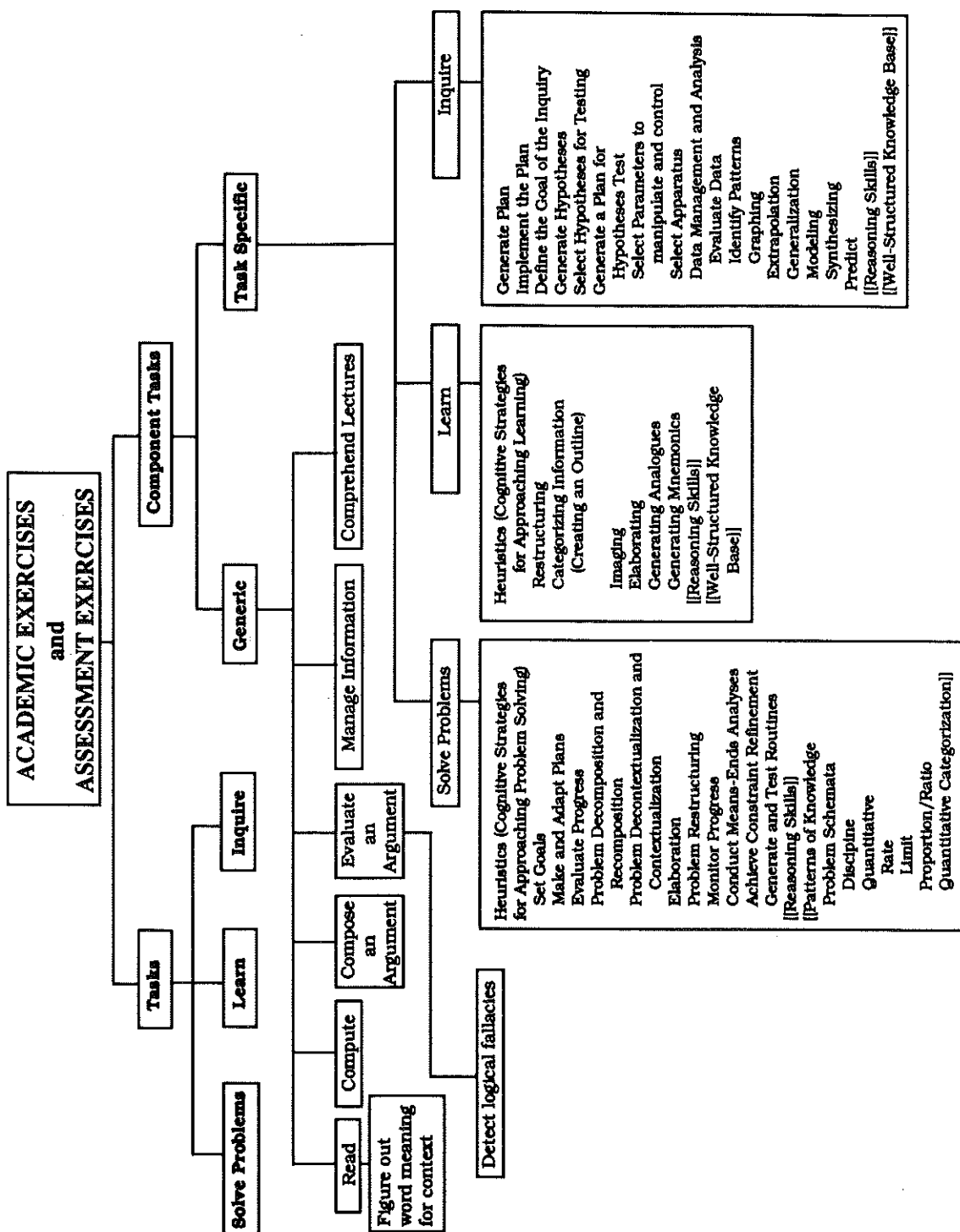


Figure 3. Taxonomy C



TAXONOMIES OF THINKING SKILLS

Differences in the structures of Taxonomies A, B, and C signal several important characteristics of thinking skills and theoretical issues related to them. These include:

- the influence of discipline on the terminology professionals use for thinking skills;
- the distinction between higher and lower order cognitive skills;
- the distinction between generic and task or discipline specific thinking skills;
- the distinction between thinking skills and academic/assessment tasks;
- the relationship between thinking skills and academic tasks/assessment exercises;
- the distinctions among modes of thinking;
- the relationship between complex thinking processes and the component thinking skills that comprise them.

Differences in the titles of Taxonomies A and B—Cognitive Skills and Thinking Processes—illustrate the influence of discipline on terminology used for mental processes. Professionals from different disciplines use different terminology for mental processes. The terms *thinking skills*, *reasoning skills*, *intellectual skills*, *mental processes*, and *cognitive skills* are used interchangeably in the literature reviewed. Usage preferences appear to be discipline dependent. For instance, the term *cognitive skill* is generally used in the cognitive psychology literature whereas educators tend to use the term *thinking*. These differences raise questions about the relationships among the terms: Does the use of different terms indicate different meanings? My personal experience suggests that professionals typically use the terms interchangeably in conversation, show a discipline-related preference for certain terms in their professional writing, and attribute different meanings to the terms when the question of definition arises.¹

The difference in the titles of Taxonomies B and C draws attention to the blurred distinction between thinking skills and academic tasks. The literature search with which I initiated this project was for terms that define processes occurring in the mind. This is the sense in which all the terms were used.² However, a close look at Taxonomy A reveals that some of the terms identified as cognitive processes in the literature more accurately define academic tasks than processes that occur in the mind. For instance, computation, reading, and managing information are identified in

the literature as cognitive skills but, in fact, more accurately describe academic tasks, not the mental processes necessary to achieve them.

Taxonomies A and B also highlight the important distinction between higher- and lower-order thinking skills. Taxonomy A's primary division is between higher- and lower-order thinking skills. However, no lower-order skills are in the taxonomy. Candidates for lower-order skills do exist. For instance, concrete operational thinking is a possible candidate. However, the logical structures characteristic of concrete operational thinking are quite complex. Even the performance of relatively simple academic tasks (subtraction with borrowing, for instance) requires complex mental processing (Champagne and Rogalska-Saz, 1984). As more is learned about the complexity of the so-called lower-order thinking skills, the usefulness of the distinction between higher- and lower-order thinking skills comes into question. Even so, the nature of the distinction should be explored further.

Is the distinction:

- a matter of *age*—is children's reasoning lower order and adult's higher order?
- a matter of experience or formal education—is an automotive mechanic's reasoning lower order and a mechanical engineer's reasoning higher order?
- a matter of difficulty in teaching or learning—can lower-order skills be taught easily to all but higher-order ones learned by only a small portion of the population?
- a matter of the frequency of their occurrence in the general population—are certain reasoning skills higher order because only a small proportion of the population exhibits them?
- a matter of performance facility—is a skill lower order if it can be performed quickly and unconsciously?
- a matter of complexity—are higher-order skills simply a concatenation of lower-order skills or are they qualitatively different (as in concrete operational and formal operational thinking)?

Taxonomy B is consistent with the view that the more formal modes of thought are concatenations of higher-order thinking skills that are of lesser complexity than formal modes of thought. It still leaves unanswered questions about the relationship

of formal modes of thought and higher-order thinking, however. Are formal modes of thought simply concatenations of higher-order skills or are they mental skills of an entirely different nature, and, if so, what is the nature of the difference?

Taxonomy B makes a useful distinction between thinking skills that develop naturally and those that develop as a result of formal education. Thus, it provides a rational way around the vacuous statement that schools must teach students to *think*. Rather, it emphasizes that schooling should change the quality of the thinking. The taxonomy, which contains ten thinking modes, reflects the nature of the change. The "ordinary" mode was invented to make the important distinction between thinking that has not been significantly influenced by deliberate attempts to make it more logical, scientific, rational, quantitative, or reflective.

Taxonomy B simply lists modes of thought without further categorization. Several possibilities for further categorization are possible. One is to make *ordinary* and *formal* new categories. Logical, mathematical, and scientific modes of thought are easily categorized as formal. However, this system presents a dilemma. Rational, creative, qualitative, and reflective thinking are characteristic of both ordinary and formal modes of thinking. Should there be another superior category or does the dilemma imply that the ordinary-formal distinction is not a useful one? Other structural organizations suggest a variety of interesting relationships among the modes of thought. They might be ranked according to some other criteria—attainability, complexity, or value, for instance.

Another possibility is to develop a categorization that subordinates other modes of thought to scientific thinking. Scientific thinking is not only rational, logical, mathematical, and quantitative but also qualitative, creative, and reflective. This line of thinking also suggests questions about the relationships between mathematical and scientific modes of thought: What features of scientific reasoning overlap with those of mathematical reasoning? Is scientific reasoning a component of mathematical reasoning? Is mathematical reasoning a component of scientific reasoning? How do either of these modes of thinking relate to desirable thinking in the humanities or engineering or law?

Searching for a plausible relationship between other modes of reasoning and scientific and mathematical reasoning is one way to approach categorization; another is to consider the overlap of the component skills of science and mathematical modes of thought. Presumably, the overlap encompasses component skills that are generally applicable to both math and science.

Another possible categorization for modes of thinking depends on whether the modes are deemed content specific or generic. This division is a major one in Taxonomy A and illuminates a central issue in psychological research and in educational practice: How does knowledge about a discipline or context influence thinking skills? Researchers in several fields have demonstrated the facilitating influence of context on the application of thinking skills. This work provides empirical evidence that performance is better on tasks presented in familiar contexts than on tasks that require the same logical processes presented in abstract or unfamiliar contexts. For instance, a person may be able to reason logically to solve a problem posed in the context of repairing an automobile but not to solve a problem requiring the same reasoning skills posed in the context of cooking (see Heller, Ahlgren, Post, Behr, and Lesh, in press).

The three taxonomies and the issues they raise are presented to support the contention that taxonomy development can contribute to attaining better theoretical understanding of the relationships among thinking skills and consequently better definition of them. Neither the taxonomies nor the issues presented above are meant to be definitive. Rather, the conclusion is that taxonomy development is a strategy that might contribute to improved understanding of the nature of thinking skills.

OPERATIONALLY DEFINING THINKING SKILLS

Advancement in teaching and assessing science-related thinking skills is contingent not only on defining relationships among the skills but also on theory that relates performance on academic tasks and assessment exercises with the mental processes that produce the performance. Whether the task is to design exercises to assess targeted skills, to categorize assessment exercises according to the thinking skills required to perform them correctly, or to make inferences about mental processes from performance on an exercise, expert consensus that the task has been accomplished is difficult to achieve. This section of the paper explores why consensus is difficult to achieve and proposes a strategy for developing operational definitions for thinking skills that explicitly relate assessment tasks and assumptions about the mental processes that are necessary for their successful performance.

The single most important factor contributing to lack of consensus is that many plausible models of mental processes can predict and explain performance on assessment exercises. A consequence is that experts design, categorize, and interpret

performance on the basis of private models. Because private models differ, their assumptions about the cognitive demands of assessment exercises and interpretation of performance differ.

Plausible and Private Models

The potential for several plausible models for any task is illustrated by the two science-assessment exercises that follow. The subject matter and purpose of the exercises are the same, but the task is different. One exercise requires students to observe and interpret a science demonstration. The other requires them to select the correct answer on a multiple-choice item. Both exercises assess thinking skills. An example of the observe-and-interpret type of exercise involves a student observing glass cylinders containing different quantities of sand rolling down an inclined plane—one cylinder is empty, one half-filled with sand, and the third filled. The student is shown a fourth cylinder one-quarter filled with sand and asked to predict the cylinder's motion down the plane and to explain the reasoning behind the prediction. After observing the motion of the fourth cylinder, the student is asked whether the observation matches the prediction and, if not, to explain why not. In contrast, a short answer exercise presents a written description of the demonstration using text and diagrams. The task is to select one of five choices that best describes the motion down the plane of the quarter-filled cylinder.³ The intent of each of these exercises is to assess thinking skills.

Exercises to assess thinking skills are generally more difficult to interpret than those designed to assess factual knowledge.⁴ A correct response on an exercise that assesses factual knowledge indicates that either the individual knew the information or was able to figure it out using information in the stem. The purpose of this type of task is not to assess the thinking skills necessary to comprehend information in the stem, to retrieve applicable scientific information from memory, to reason from the information in the stem to the correct answer, or to eliminate incorrect responses. Neither is the intent to assess the mental strategy the individual used to reach the conclusion. The intent is only to ascertain whether or not the individual can match the required information with the correct answer. In contrast, the intent of exercises such as the rolling cylinder is to assess not the accuracy of the answer but the mental strategy that generated it.

In practice, however, it is assumed that a scientifically correct answer was based on higher-order thinking. This is not necessarily the case. Several plausible explanations can be posited

that will produce a scientifically correct response to an exercise meant to assess thinking skills. For example, if the exercise is to explain a physical event (the reasons for a cylinder's motion down an incline), a correct explanation may result from being familiar with similar events and remembering the explanation for them. For an individual with the relevant experience, the exercise is, in fact, a knowledge recall item. Alternatively, the person may be unfamiliar with the situation but recognize that a scientific principle he or she knows is applicable to the situation and reason with the information presented and the physical principle to the correct answer. Another possibility is that the person uses a wrong assumption or incorrect information, or reasons illogically, but still arrives at the correct answer.

It is especially important to realize that an incorrect answer may signal misinformation, not flawed thinking. A person might make an incorrect factual assumption in applying a correct scientific principle or rules of logic and produce an incorrect answer. Since the answer depends both on information and processing that information, there is no way, on the basis of the response alone, that the observer can know whether the answer selected is the result of recall, logical thinking with correct information, flawed thinking with wrong information, or logical thinking with wrong information.

Because exercises of the observe-and-explain type yield information about the thinking that contributed to the answer, results from the administration of this kind of exercise provide data that can be used to choose among alternative interpretations.

Further Confounding Factors

In addition to uncertainty about the correct model of mental processes, other factors that confound interpretation include the quantity of data stimulated by exercises, age, and the age-related correlates level of cognitive development and breadth of experience.

Before observers can reach consensus on interpretation of performance, they must agree on their observations. Although the assumption is that agreement on observations should be easily attained, this is not always the case. The complexity of the performance elicited by the task influences how well different observers will agree on what the students have actually done. For instance, student performance on the exercises described above is more complex in the task that uses real objects than in the paper-and-pencil version. The first task elicits both student actions with

the materials and extensive verbal output. The large quantity of data generated presents both advantages and disadvantages. It makes agreement among observers more difficult to achieve, but provides more data on which to base interpretations. More data requires observers to agree on principles for determining which data are relevant. In addition, the quantity and richness of the data make the arguments linking data to performance highly involved. As a consequence, assessment of complex performance is more influenced by observers' models of thinking processes than is assessment of short-answer exercises.

In the case of short-answer exercises, agreement on the observation is easily obtained. Either the student chooses the correct description of the cylinder's motion or does not. However, reaching consensus on the mental processes that can be inferred from the correct response is as difficult as for the observe-and-explain exercise. An exercise that requires complex thinking for a typical eight-year-old is often simply information recall for the typical 12-year-old. Thus age and experience must be factored into interpretation of performance.

Developing Empirically Based Models for Mental Processes

Developing theory that links assessment exercises and mental processes involves observing students in the target age group as they describe the thought processes they use to generate answers to an exercise. On the basis of such protocol data, models of the knowledge and mental processes used by typical students can be developed and exercises can be categorized on the basis of the most frequently used model.

The methodology for gathering the empirical data and developing the theory necessary to reduce the ambiguity of inferences made about thinking skill level from performance have been developed by cognitive scientists. They use this methodology to develop models of cognitive processes that underlie the performance of complex tasks such as playing chess, diagnosing disease, and prospecting for oil. Applying this methodology to assessment exercise development would involve collecting protocol data from students and, on the basis of these data, developing models of the information and information processing skills used by typical students.

This methodology has its critics. The critics raise significant conceptual issues related to the validity of the models. For instance, there is no way to be certain that a verbal protocol is a true reflection of how an answer was arrived at rather than an

after-the-fact fabrication of how it might reasonably have been arrived at.

This research poses intellectual challenges. In addition, its results may challenge the basic assumptions of current assessment strategies. It is possible that this research will show that students in the same population use different strategies. This finding would contradict an assumption of current assessment procedures, namely that processing strategies are consistent among individuals in a defined age group.

Evidence exists suggesting that the consistency assumption may be invalid. For instance, one researcher reports that on a highly abstract spatial reasoning task frequently used in instruments to measure intellectual ability, adult subjects use at least two distinct processing strategies (Embretson, 1988). Embretson's goal is to apply cognitive design principles to psychological testing. Her method involves developing models of the mental processes applied in the solution of a spatial reasoning task. The task is to compare a two-dimensional figure with four three-dimensional figures to determine which of them represents a possible folding of the two-dimensional figure. Embretson reports that she has identified at least two strategies individuals employ to perform this highly abstract task.

Embretson's work illustrates the difficulties that arise in applying cognitive theory to developing assessment exercises. Developing taxonomies and performance models are complementary processes. Taxonomies help define what is to be taught and assessed. Operational definitions describe procedures for their measurement. Progress in teaching and assessment of thinking skills requires extensive work to develop a taxonomy that serves the purpose of thinking-skill definition. Achieving this goal is a resource-intensive but necessary initiative.

CONCLUSION

Existing assessment and teaching strategies are inadequate for testing and teaching thinking skills. It is misguided policy to continue expending vast resources collecting data with assessment exercises of questionable validity. Instituting a research and development program is a necessary first step toward developing valid assessment instruments for thinking skills. High on the agenda of such an R&D program must be to develop definitions of thinking skills. The R&D program necessary to develop the empirical and theoretical basis for research-based teaching and assessment will require considerable resources. The exercise development strategy

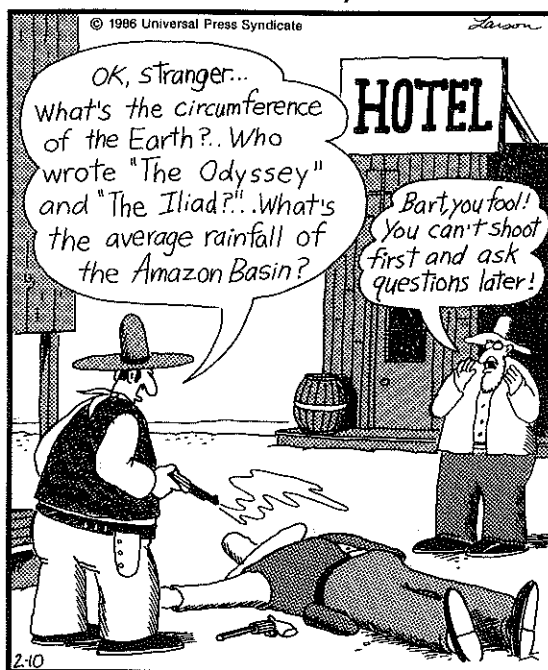
proposed is labor intensive and expensive to initiate. However, to continue using data of questionable validity to formulate policy and guide classroom practice is irresponsible.

Validity of Science Assessments

Jerome Pine

THE FAR SIDE

By GARY LARSON



THE FAR SIDE COPYRIGHT 1986 UNIVERSAL PRESS SYNDICATE.

Reprinted with permission. All rights reserved.

INTRODUCTION

Is a particular science assessment valid? Does it meet a dictionary definition of being "sound, based on premises which imply the conclusion"? Does a positive test result correlate with competence as measured more fully in nontest situations? The dictionary defines validity as the "quality of being valid, as 'to question the validity of an intelligence test'." Clearly, test validity is not a novel idea! Nor is it an esoteric one. Yet it is ignored entirely in the reporting of standardized test results in the media and in reports to policy makers. The reason, I believe, is that, though the question—Is an assessment valid?—is simple, the answer is not. However, the need to establish validity is so basic that it cannot be ignored. Assessments that are not validated, yet are taken seriously, can mislead and can harm the cause of sound science education.

Jerome Pine teaches and does research in physics and neuroscience at California Institute of Technology. He has also developed materials for hands-on elementary-school science. He is currently involved in a study of science testing methods at grade 5.

This discussion will be limited to tests, the kind of assessment that now dominates. But it includes tests in the broadest sense, including, for example, those that involve performance of investigations as well as free-response paper-and-pencil items. A test, as defined here, is any attempt to judge a student's science knowledge and competence from a sample of work done in a short period of time (perhaps one-half to a few hours), as contrasted to longer term (and potentially deeper) assessments such as a teacher's judgments based on long-term knowledge of a student, or the evaluation of a portfolio of a student's work.

A test must be valid in several respects, each of which poses its own particular requirements:

1. *Valid content.* A test must be based on, and reflect, a curriculum in the broadest sense—not just in content and concepts but also in competencies related to pursuing a scientific inquiry. The curriculum may not be a particular textbook curriculum or “taught curriculum,” but must reflect the knowledge of science and the doing of science that the testmaker, *and the test user*, accept as a standard of science competence.

2. *Valid methodology.* Is performance on the test a valid measure of a student's mastery of the body of knowledge to be tested? This is the most difficult requirement for test validity, and one that is seldom, if ever, addressed. It embodies the ideas of “construct validity” in the literature of assessment, characterized by Messick in a recent paper as “... a sine qua non in the validation not only of test interpretation but also of test use...” (Messick, 1989). For example, do multiple-choice questions that are categorized as testing science process skills or “higher-order thinking skills” actually measure those abilities? An important aspect of this question is whether the test is free of biases that have nothing to do with science knowledge but arise from insensitivity to the effects of the students' sex, cultural background, linguistic ability, and socio-economic level.

Another aspect of methodology is the statistical machinery for dealing with questions of reliability, reproducibility, comparisons among different versions of a test, and the like. Effective methods for meeting these needs have evolved over years of standardized testgiving, and this area of methodology will not be addressed here.

3. *Valid reporting.* Does the reporting of test results adequately describe the knowledge and abilities that the test is intended to measure, and does the report provide evidence that the test is in fact a valid measure of that body of knowledge and abilities? Are test results presented in such a format that results for different aspects of science knowledge can be seen separately? If a

single total score is presented, is the relative weight of the separate aspects of science knowledge apparent? Is a measure of the statistical uncertainty of the score or scores provided?

This paper discusses these validity areas, emphasizing problems and possible solutions as they relate to tests that evaluate populations of children rather than individuals. Such tests are being used more and more to determine whether science is being taught well at the national, state, district, or school level. For such tests, valid methodology need be considered only for the aggregate of students. The requirement that a test accurately reflect the knowledge of any and all the individual students poses a much more difficult validity problem. The range of individual differences is so large that sole reliance on any test is inadvisable and almost certain to be flawed. In England, for the new tests to begin in 1990, long-term classroom assessment and teacher judgment will be used in conjunction with tests that are not multiple choice (Task Group on Assessment and Testing, 1988). However, in the U.S., multiple-choice tests are used when individual testing is required for various federal support programs. The results become part of a student's record and are reported to parents.

A few examples of test items are used in the discussion below, not to pass judgment on the tests they come from but to illustrate questions of valid content or valid methodology. The examples are characteristic and exemplify issues that relate to current widely used tests.

CONTENT VALIDITY

The testmaker must have a clear view of the body of knowledge being tested, and this must be made explicit to the user. At this time, the content of elementary school science exhibits enormous variation among classrooms and schools and across states. Textbooks display general consistency and a commitment to superficial treatment of a very broad range of factual knowledge, with few hands-on activities for the students. However, the amount of nontext-based teaching is sizable and growing. Here, emphasis is on selected subject matter, treated in depth, with hands-on student involvement. The two extremes are not only very different but also incompatible. And there is a wide middle ground of partial adherence to each approach. Given this situation, the content validity of tests cannot be taken for granted. In particular, tests that are dominated by questions that assess rote learning of textbook facts are ill-matched to assessing inquiry-based, hands-on learning.

If the test user finds that the test's content is in harmony with his or her curriculum, then one necessary condition has been met. If, however, the educational aims of the user and the test are

mismatched, the test is not useful. In the best of all worlds, the dissonance of a known mismatch will have a constructive effect. If the test is good and the user's curriculum is poor, perhaps the mismatch may even exert a positive influence. If the converse is true, and, in the worst case, the test is mandated by a higher authority, making the conflicting curricula explicit offers a possibility for negotiating a change, or at least points up the problem of interpreting the test results.

The case of a school or district that offers an exemplary hands-on science program along with a state- or school board-mandated standardized test of the usual types provides a concrete example. If the publisher explicitly specified that the test largely measures common sense, plus rote-learning of text-book knowledge common to those texts widely used in the U. S., the mismatch would be clear. However, a test publisher is very unlikely to characterize its test in this way. More likely, the publisher would characterize the test as a mixture of those basic concepts and processes essential to a sound knowledge of science. There is a problem here, and often not a subtle one. The mismatches can be enormous.

Too often, items from widely used standardized tests assess knowledge of neither fundamental science concepts nor process skills. They are a form of IQ test, free of science content. Any relation between the curriculum on which this test is based and hands-on, inquiry-based, science is coincidental. The content of these tests is usually closely guarded by the publishers; the tests are not available to a teacher, a parent, or a legislator—except, perhaps, through a few “samples” not actually used on the test.

Publishing the tests so that the items would speak for themselves would be a very important aid in establishing valid content. So, too, would establishing an independent review board to examine and characterize tests. Such a group would need to be a carefully selected mix of scientists and science educators, constituted by a recognized authority, such as the National Academy of Sciences or the National Science Teachers Association. Users could then insist on tests that the board had characterized. A publisher could hardly claim that the need for secrecy precluded study of a test by such a board. The existence of a review process like this would, of itself, exert a strong influence on publishers to characterize their products more objectively from the start.

A powerful tactic is for a state to outlaw use of a test unless the publisher agrees to make it public after being given, as New York did successfully with the Scholastic Aptitude Test. Some have suggested that this could raise test-making costs prohibitively, but the cost to students taking the tests in New York is only

slightly higher than in the rest of the nation. As a result, "...the College Board makes public five editions of the SAT each year as part of its ongoing program to provide full public information about these tests." (College Entrance Examination Board, 1988). A quite different, and discouraging, use of legislative authority is the Congressional edict that the National Assessment of Educational Progress may not release items used for trend data for 10 years after their first use.

Creating a review board for tests is not a simple matter, since the evaluation of test items against the aims of a given curricular standard will be done by opinion, and individual reviewers are bound to exhibit differences. As an example, during a recent indicators study (Murnane and Raizen, 1988), a panel of 10 scientists and science teachers, selected for their expertise in science and science education, reviewed the content of nine leading standardized tests. The group shared a commitment to teaching science process skills and fundamental concepts rather than a volume of factual detail. Yet the variations among reviewers were so large that no sound conclusions could be drawn about the relative merits of the tests. The group did conclude that the tests on the whole were poor matches to a process-oriented science curriculum.

VALID METHODOLOGY

Even when a test's subject matter is matched to a desired curriculum, determining whether the test is a good measure of knowledge of that subject matter is still not easy. A test question may fail to succeed in a variety of ways. It may be ambiguous; it may depend on nonscience skills or knowledge; it may be a pencil-and-paper surrogate for knowledge of a science process that misses its mark; and so on. One particularly severe problem with multiple-choice questions is whether they measure science knowledge, general knowledge, or skill in ruling out multiple-choice distractors. For example:

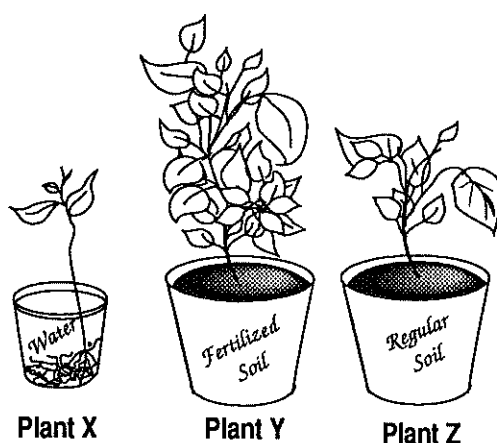
	A light-year is:
80.2%	A. the distance light travels in one year
9.6%	B. about one million kilometers
5.7%	C. a parsec
3.4%	D. an angstrom

What answer would a bright kid who never studied astronomy give? Is this a good astronomy subject-matter measure, as it is supposed to be on the California Assessment Program (CAP) eighth grade science test? It is reprinted in a report, with the comment that students scored higher on it than on any other question

about astronomy, and no further analysis (California Assessment Program, 1986).

This CAP test is new, and in many ways seems better than most multiple choice tests. The questions are categorized into subject matter areas, science and society, manipulative laboratory skills, and science processes. They are designed to match a state curriculum that emphasizes a hands-on approach to teaching science and process skills. Nonetheless, the validity of many CAP questions seems doubtful. In science process areas, multiple-choice questions seem particularly inadequate. Figure 1 shows two questions that are supposed to assess a student's ability to draw simple inferences.

FIGURE 1. Questions Designed to Assess a Student's Ability to Draw Simple Inferences



Which one of the following conclusions is best supported by the above illustration?

- | | |
|-------|---|
| 1.1% | A. Plants grow better in water. |
| 2.2% | B. Plants grow better in regular soil. |
| 1.3% | C. Plants grow better in glass pots. |
| 93.4% | D. Plants grow better in fertilized soil. |

Recently, some forests were cleared in the Himalayan Mountains. What conditions would most likely occur as a result of this clearing?

- | | |
|-------|--|
| 27.1% | A. Colder weather would occur in the hills. |
| 14.9% | B. Less rain would fall on the plains below. |
| 11.2% | C. Snow would fall in the mountains. |
| 44.8% | D. Floods would occur on the plains below. |

The first question is so trivial as to be a-scientific. Given the three pictures, students need no inference-forming ability to pick the big healthy plant. Can this question in any way indicate learning of how to form scientific inferences reliably? The second demands an answer based on recall of rote knowledge. Again, inference-forming skill, based on the weight of evidence, the control of variables, etc. is irrelevant to answering the question.

Though process without content is probably not a feasible ideal, questions categorized as process-measuring, but which depend entirely on content, are clearly misrepresented. An example from the 1986 NAEP released items:

Level 300 - Analyzes Scientific Procedures and Data

Which of the following best explains why marine algae are most often restricted to the top 100 meters of the ocean?

- A. They have no roots to anchor them to the ocean floor.
- B. They are photosynthetic and can live only where there is light.
- C. The pressure is too great for them to survive below 100 meters.
- D. The temperature of the top 100 meters in the ocean is ideal for them.

These examples clearly show that at some level it is possible, and probably practical, to judge validity of methodology by opinion, just as content may be judged. Of course the same difficulty of dealing with the variation of opinions among the members of a review body must be met.

If a reviewer judges methodological validity and finds no obvious flaw in a question, only a necessary, but not sufficient, condition is met. A question may look good but still not work the way we think it should. Anyone who has made up and graded exams for students he or she knows well must be aware of how easy it is to generate an invalid question without realizing it, through difficult or misleading language or poor organization. The question in figure 2, from the British APU (Assessment of Performance Unit, 1985), might look good but poses some potential problems.

This APU question is designed to test the student's ability to do an experiment. But does it? We need to ask:

1. Is the question too linguistically demanding for 11-year-old native English speakers to understand?
2. Is it too linguistically demanding for nonnative speakers?

FIGURE 2. Assessment Question from the Assessment of Performance Unit

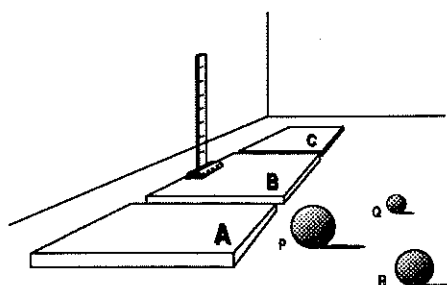
You have probably noticed that some balls don't bounce as well on some surfaces (like carpet) as they do on others (like tarmac).

Suppose you have 3 different balls, P, Q and R, and 3 different kinds of surface, A, B and C.

You can also use a long ruler in a stand. You can use any of the things in the pictures below (but you don't have to use all of them).

Write down what you would do to find out:

If the ball which bounces best on one surface also bounces the best on all the surfaces.



Make sure you say:

- Which things you would use
- What you would do
- How you would find out the results

3. Is it conceptually confusing, or so dense with information, that only very bright kids will be able to handle it?
4. Is the requirement to write the free-response answer so demanding that the question tests writing as much as science?
5. Does writing about doing an experiment equal doing it?

You can probably raise other questions. Can written questions that attempt to create a genuine challenge to process skills, and thereby test for them, clearly avoid such problems? It seems problematic. How can we resolve such issues and decide whether a question is flawed or valid? My examples moved from the most trivial multiple-choice questions to a very complex free-response

problem, but, of course, a range of intermediate methodologies exists. These also pose this sort of validity problem, even if they are free of the most basic objections.

The British Assessment of Performance Unit is most likely the world's most experienced group in developing test questions for assessing science process skills. Given the task in 1977 of assessing science knowledge in England, Wales, and Northern Ireland, the leaders (who had developed hands-on elementary school science curricula in preceding years) rejected existing tests and constructed highly process-oriented examinations. Their tests, developed and used over a 10 year period, even include the observation of children performing investigations with real apparatus. Yet the APU does not have data to establish the validity of its questions. Except by judgment (prejudice?), we do not even know whether good performance on observed investigations correlates with effective hands-on science education.

Validating methodology is an area in which informed opinion will not suffice. Can we find some objective way to establish whether a question measures what we hope it does—e.g., whether the APU example measures the ability to do a meaningful experiment and to interpret the results correctly? I believe there is a way worth investigating, but it remains to be seen whether it is practical. In principle, the idea is simple: Use an assessment that is more valid than the test, though it may be too difficult to use as a general method, to calibrate the ability of the test questions to give a valid measure. An obviously valid assessment is the detailed judgment of a skilled teacher who has worked with a child in hands-on science for a school year.

An alternate approach, which is not based on judging individual children's science knowledge, is to try out questions on matched populations of children who differ in the average quality of their science education. For example, children from schools where the science curriculum is hands-on and inquiry-based might be compared with those from schools with little science education or with predominantly text-book based education. If the two populations are well matched and each has the same variations in general knowledge, cognitive ability, cultural background, and linguistic ability, the effect of science education independent of other factors should be visible. If other factors distort the result, that can also be seen. Of course, such a comparison does not prove that a question actually measures the process skills we may think it does, only that kids who have experience in doing and interpreting experiments do or do not show better ability to answer a given question correctly.

Rich Shavelson and I are beginning a study that will use observed investigations as a benchmark assessment to validate less

difficult and expensive methods. This study will also use children from matched populations who come from schools with differing levels of science instruction.

In summary, some progress in establishing methodological validity can be made by informed judgment, but this seems more likely to eliminate poor questions than to establish the usefulness of potentially good ones. We need objective techniques. There are possibilities based on comparing test questions with more valid indicators of science knowledge but still no easy answers.

VALID DATA PRESENTATION

As mentioned above, any test result must be interpreted in the context of the curriculum or set of valued goals it claims to match. Furthermore, if the test is believed to be methodologically valid, the grounds for this validation should be stated. For example, the score on the NAEP science test is widely used to indicate the state of science education across the country. Small changes in a single overall score are given wide publicity and used to argue for policy changes. The methodological validity is neither questioned nor discussed in the reports that are widely disseminated.

I analyzed 100 items of the 1977 NAEP test for content relevant to an inquiry-based curriculum that emphasized fundamental concepts but not rote knowledge of details. In my opinion, one-third called for pure rote recall of detail, one-third were adequate, and only one-third were good. A decrease in the aggregate score on such a test might mean worse scores on the rote detail questions because students spent more time learning how to make investigations. Or it may mean less fundamental conceptual knowledge. How are we to know? What does the change mean?

A good science curriculum, and a test matched to it, is bound to have a mix of subject areas in which basic concepts are taught: life science, physical science, earth science, technology, etc. In addition, both will cover an array of process skills, such as the ability to organize observed data, infer, generalize, hypothesize, control variables in an experiment, etc. Finally, in the best of all possible worlds, they will also cover the ability to pursue an investigation—to build from content and process the ability to do science. Even if the individual scores on different aspects of such a test are combined to produce an overall score for some purposes, teachers, policy makers, and other professionals must have access to results for all the subcategories—even for the individual items if they are published. Furthermore, naming the subcategories without establishing the validity of the real test questions is misleading if not meaningless.

Is it possible to justify reporting a single assessment score on a test that covers a variety of subject matter? Maybe. Combining a great range of data, related but different, into some sort of average is common in other contexts. We see it in the newspapers every day with regard to the gross national product, the state education budget, the total number of nuclear warheads, and so on. Sometimes we do need single numbers to deal with complex aggregates. But in almost every case, the disaggregated data are available to those who need to make serious judgments. The professional economist, the legislator, and the arms control negotiator insist that it be. Furthermore, when the issues demand it, the disaggregated data do find their way into the newspapers. The professionals can bring them to the attention of the public when they believe it is necessary.

In testing, the problem is not only in the single score but also in the poor availability of the detailed data. Valid test interpretation requires different data in different contexts, and they all must be available. Finally, if an aggregate single score is derived, the weighting of subcategories must be described. For example, in the case of the California Assessment Program test, the subcategory data and weighting are both available. A school that emphasizes hands-on science can point out that only 20 percent of its aggregate score is based on science process skills. And they can publish their score on that subset of the test. If the test is valid, and their teaching is effective, they should be able to show their success in teaching what they believe is most important.

Of course, if the test itself is flawed in content or methodology, good data presentation is of no consequence. The necessary condition must be a valid test. Then, the valid presentation of data is an issue the test-using community should be able to address effectively.

SUMMARY

Establishing test validity is crucial if we are to believe—and use—the results of tests our students take. That validity is necessary in three areas: content, methodology, and data presentation. All must be considered if we are to draw valid conclusions from test results. Informed opinion can be an important tool in assessing valid content and methodology. Furthermore, the user's insistence on complete information can improve the validity of test reporting. At present, neither of these is being used effectively. As efforts are increased to establish good inquiry-based science curricula in elementary schools, tests appropriate for the children in these schools must be developed and validated, and these tools should be used.

Problems of methodological validity that cannot be resolved by opinion remain. These may be partially resolvable by experiment. In the final analysis, experiment cannot prove, in a theory-independent way, that a test question measures the particular knowledge that we think it does. However, experiment can show whether the ability to answer the question correlates with that knowledge and to what extent nonscience attributes such as cognitive ability, linguistic skill, etc., influence the correlation.

In many instances, publishers of science tests have used the correlation of scores on their tests with general academic prowess and cognitive ability as evidence for test validity (Wall, 1981). Whether the tests are successful indicators of science knowledge, rather than tests of other abilities, has been almost entirely ignored. There has been very little research in objective validation of tests as measures of science learning, as contrasted with validation by opinion. In view of the important role effective, objective methods could play, more thought is needed regarding how such validation might be achieved, and more research is needed to test existing and new ideas.

Assessing Science Education: A Case for Multiple Perspectives¹

Frank E. Davis

Imagine this situation. We have entered a fourth-grade classroom with several observers. Our objective is to assess how and what children are learning through an Elementary Science Study life-science curriculum called Eggs and Tadpoles. As we enter the classroom, we find several groups of children engaged in different activities. We observe children in one group looking at frogs' eggs in various stages of development under a microscope. In another group, we find children constructing a chart entitled "The Life Cycle of Frogs." In yet another, we find children carrying out an experiment to determine the effect of various temperatures on the growth of tadpoles. Finally, we find a fourth group of children reading materials that describe frogs as amphibians having smooth and moist skin, webbed feet, and long hind legs adapted for leaping. We also note that some children seem engaged in reflective activities, while others are actively engaged in manipulating materials; some look interested in what they are doing and some look bored.

How do we assess what is happening in this learning environment? We might hear one observer state that some learning is clearly taking place because of the way children perform certain activities—performances that should be the behavioral objectives of the science curriculum and should be rewarded and guided by the science teacher. For example, such an objective for these young children might be to carefully observe and record changes as tadpoles transform into adult frogs. Clearly, one of the goals of any science curriculum is to get children to behave like scientists and controlled observation is part of scientific experimentation. Children's performance of controlled observation can and should be assessed.

Given this behavioral goal, another, more complex, performance might involve a child's association of the many observed instances of transformations of tadpoles to frogs with a general concept of transformation represented by the word metamorphosis. Scientific knowledge as a system of concepts and relationships between concepts is fundamentally a generalization of concrete experience. This complex performance is marked by the acquisi-

Frank E. Davis is an Associate Professor of Research and Evaluation in the Division of Advanced Graduate Study and Research, Lesley College, Cambridge, Mass. He has taught at Lesley College for four years. Prior to joining the staff at Lesley College, he taught mathematics at the University of Massachusetts, Boston, for 10 years.

tion and use of a scientific language. Consequently, in a child's language behavior we can also observe the results of science learning.

In general, the first observer concludes, we can assess the adequacy of this fourth-grade classroom as a science learning environment by comparing all behaviors, including language behavior, that we expect to observe as children are guided through a concrete learning experience. In fact, the science curriculum should be defined through a set of expectations about children's behavior, and a good science teacher should be one who can create experiences that promote such behaviors. These expectations about children's behavior should reflect the basic idea that knowledge is most reliable and valid when it is generalized from a scientific experience—an experience that is created, described, and controlled by the special behavior of the scientist.

A second observer takes a different tack. She says that the point of any good science curriculum is to begin to give kids an awareness of forms and of scientific laws that explain these forms and changes in form. For example, while kids observe many different tadpoles changing into frogs, they also can abstract the general form of a species and a general principle that describes how frogs develop. Some children may go further and abstract the concept of a general form of amphibian and a general biological principle that defines the idea of a life cycle for living forms. It is this process of learning by abstract conceptualization that must be measured in the classroom. In general, scientific knowledge describes reality in terms of the most fundamental forms (the subject of physics) to the most complex forms (the subject of biology). Even though the knowledge of such forms may be constrained by the way human beings can experience reality, a scientist aims for knowledge that is universal.

Further, this classroom observer comments that science experiments should be used to create an environment in which children can practice abstract thinking—an environment in which rational thinking becomes scientific thinking. A good scientist is able to deduce as well as induce knowledge, producing a system of scientific knowledge that is rational and appears to stand above concrete experience.

Clearly, this observer concludes, what should be assessed in this classroom is the scientific concepts and relationships between these concepts that kids obtain. The fourth-grade science curriculum should have a stated standard of acquired scientific knowledge that can be tested. Equally important, curriculum must be designed to engage children in rational and abstract thought.

Also, assessments must be made with the understanding that not all kids are alike—not all are equally intelligent or capable of scientific thinking.

A third observer enters the discussion. From his perspective, we must understand the complex interaction between what children believe and how they see reality before they begin the curriculum and the more adult view of reality that the curriculum is designed to produce. The dissonance created in the child's mind because of his or her activity in the classroom is important. For example, if a child believes that a frog always looks the same, changing only in size throughout a life cycle, then the observation of physical changes in form as a tadpole becomes a frog creates a dissonance in the child's thought. This dissonance creates an opportunity for learning because it forces children to reconsider what they have already learned and to change their ideas to better represent reality. This learning involves reflective abstraction. In fact, it is this process over a collective human history that characterizes the history of scientific knowledge. The frontiers of scientific knowledge are defined by dissonance and inconsistency, and progress is measured by the construction of more consistent knowledge. Collective scientific learning is both reflective, in the context of a developing system of knowledge, and abstract in the context of describing something that frames human experience.

In addition, in a classroom of young children, this type of reflective learning is an opportunity for broadening not only knowledge but also the thought process itself. Reflective abstraction is a type of thinking that must develop. Many children in the fourth-grade still use concrete thinking—their cognitive stage can be characterized as concrete operational thinking. In this context, a goal of the science curriculum may be to release thought from concrete experiences. For example, children who are required to imagine a scientific experiment and its possible results before carrying it out may be forced to develop a new way of thinking—a type of formal or hypothetical thinking that is clearly a tool of the scientist.

In this context, the third observer summarizes, what we need to assess is the developmental appropriateness of the science curriculum. Does it provide the opportunity for either expansion of knowledge—what we might call a child's history of science—or expansion of the thought process, which might be developmentally defined as a child's movement toward abstract and reflective thought? The science curriculum should produce a classroom that facilitates change in both these areas.

Finally, a fourth observer responds to the question of how to assess science learning. This observer reminds us that a science curriculum defines a problem-solving environment. A child's activity in such an environment defines an educational context—a context in which learning occurs. For example, when kids are asked to determine the best ecological environment for the growth of tadpoles, they are engaged in a scientific problem. For children to understand this problem, they must absorb a history of human problem-solving activity that has already resulted in concepts such as ecology. Children must also develop a capacity to mentally represent the problem and possible solutions. If an educational context is understood to be part of a broader learning environment defined by a child's ongoing life experiences, the development of these capacities might be associated with a natural striving for competence.

However, the presence of a problem and the capacity to represent it are only half the context of a science curriculum. The child's activity in understanding and solving the problem is the other half. It is this natural instrumental activity that both expands the child's capacity to represent problems internally, to absorb the products of a history of human problem solving, and to define and understand new contexts for learning. The activities of understanding the problem and solving it are the activities on which we must focus.

This fourth observer notes that although scientific knowledge can appear to be abstracted into a system of concepts and relationships between concepts, this knowledge has meaning only in a context defined by human activity—activity that is bound up with human intentions and purposes. The objectives of the science curriculum and activity of a science teacher in a fourth-grade class must lead to a child who produces a context—a context that defines and allows both a scientific problem and its solution. We must define science learning as learning aimed at solving problems—problems in an environment defined and changed by human activity. To assess this learning, we must look at how well a child can define and understand a science problem as well as how well a child can solve scientific problems.

I hope the reader, upon reflecting on these imagined comments, has two impressions. On the one hand, I hope that all of these ideas about assessment seem reasonable. The ideas of assessing a behavioral repertoire, a state of conceptual knowledge, a stage of knowledge that has a developmental history, or knowledge that evolves from the ability to produce and solve problems should all have commonsense appeal. On the other hand, I hope that these

ideas, although they all seem reasonable, also seem potentially contradictory. For example, it should not be hard to imagine that one observer would disagree with the idea that scientific knowledge is primarily acquired through abstract thought without regard to concrete experience or that another observer would disagree with the idea that dissonant or contradictory concepts may be signs of impending cognitive development. In general, I have tried to craft these comments so that they sound reasonable and familiar, but yet could be the grounds for disagreement and debate.

The objective of this paper is to illustrate that there are potentially contradictory views of what should be assessed in science learning—contradictory views that may all have important implications for assessment and may not simply disappear with a more inclusive or general theory. What underlies differences in conceptions of what should be assessed in a science learning environment? Is it a conception of how science learning occurs or of how children learn? Is it a conception of what is or should be learned in a science curriculum? I have found that a useful answer to these questions lies within a much broader framework. This framework emerges from questions about the assumptions underlying any conception either of knowledge or, reflexively, of a process of knowing. This framework emerges from philosophical questions about epistemology and metaphysical assumptions about the nature of reality.

THE THEORY OF WORLD HYPOTHESES

A more theoretical debate about the origin and structure of all knowledge should parallel a debate about what is learned in a science learning environment. It is this broader philosophical debate that Stephen Pepper attempted to clarify through what he called a theory of world hypotheses (Pepper, 1941). Pepper characterized world hypotheses as hypotheses that capture a history of philosophical thought about knowledge. These hypotheses are not simply a system for categorizing philosophical arguments about knowledge. They are also products of an ongoing process of acquisition of knowledge and, consequently, are still subject to elaboration and proof.

Pepper defined knowledge as a system of beliefs ranging from commonsense knowledge to highly refined knowledge. This range of knowledge is a consequence of different types of evidence. Uncriticized evidence corresponds to commonsense knowledge; criticized evidence corresponds to refined knowledge. In Pepper's view, this range of evidence implies a process of knowing that seeks to move from uncriticized to criticized evidence. A

process that leads to criticized evidence must involve two types of corroboration: structural corroboration, which is the corroboration of "fact" with "fact" (although this does not imply that a "fact" can be identified), and multiplicative corroboration, which is the corroboration of an individual's ideas with other individuals' ideas.

Through structural corroboration, Pepper asserted, all knowledge must be produced within a relational structure in which one piece of evidence is tied to another. For example, children in a fourth-grade classroom who are asked to manipulate the environment in which plants develop to see varied effects are being asked to understand a complex relational structure. Possibly, these children will end up understanding that the chemical structure of the environment affects the developing organic structure of plants or that an organic structure is defined in terms of the functions it performs to sustain life. In Pepper's view, if we examined this knowledge carefully, we would find that ideas about chemical structure of an environment and the organic structure of a plant acquire status as knowledge only through the elaboration of some type of relationship; through structural corroboration. In this example, we should also understand that a large amount of structural corroboration must occur before such things as a chemical or organic structure can be conceptualized. This is what makes it difficult to understand what may really be a pure "fact."

Relational structures give knowledge cohesiveness. Different types of relational structures result in exclusive systems of knowledge that Pepper called world hypotheses. How does a relational structure develop that supports structural corroboration? Pepper proposed that this development can be represented by the construction of metaphors; that is, one meaning or structure is applied to another by analogy. Structures built by analogy underlie much of what we call factual knowledge. For each world hypothesis, Pepper identified a root metaphor that he believed captures its historical and continuing development.

Occurring simultaneously with a process of structural corroboration is a process of multiplicative corroboration. Through multiplicative corroboration, the corroboration of one individual's ideas with another's, all knowledge acquires an interpersonal dimension. In essence, all knowledge must have the possibility of common agreement. For example, whatever effect a child observes or understands in manipulating the environment in which a plant develops must also be observable and understandable by another child. The possibility of multiplicative corroboration is as much a part of the possibility of criticizing evidence as is structural corroboration.

Consequently, when we look at a history of knowledge we observe the interplay of structural and multiplicative corroboration. In Pepper's view, we observe a process by which greater and greater structural webs of evidence are constructed. These webs explain a collective reality with increasing precision and scope. The largest of these structures are world hypotheses. These hypotheses propose to explain the structure of all knowledge. Pepper proposed that four hypotheses have reached the status of world hypotheses: formism, mechanism, contextualism, and organicism.

Pepper's theory is valid only if it can frame both a history of knowledge and contemporary theories and debates about knowledge. In a narrower sense, it should frame theories and debates about science education and how we assess science learning. The overall test of the validity of Pepper's work is beyond the scope of this paper. Instead, I will ask the reader to suspend this broad assessment of Pepper's work in favor of analyzing its ability to order debates about science learning. In this context, I will first summarize Pepper's four world hypotheses and their implications as frames for theories about assessment of science learning. Second, I will suggest a theory that accounts for these multiple perspectives and consequently implies that a variety of assessment tools are required to evaluate science learning effectively. I begin with a description of the world hypothesis of formism.

Formism

The world hypothesis of formism portrays a metaphysical world of universal forms. These ideal forms are described by qualities and relationships between qualities. Ultimately, knowledge is an understanding of reality through form and transformations of form. The metaphysical nature of formism arises in explanations of the relationship between universal and ideal forms and a concrete reality of particulars and imperfect forms. For example, some formists say that forms in the concrete world are a consequence of the participation of qualities and relations in an underlying reality of matter; others might talk about the exemplification of the ideal in reality. The metaphor that captures the structure of this hypothesis is the metaphor of similarity. Whenever we see similar things, we are viewing the potential existence of ideal and universal forms defined by qualities and relations.

In a historical context, the world hypothesis of formism subsumes the work of philosophers such as Plato and Kant and, in a contemporary setting, the early philosophical work of Bertrand Russell. In the world of physics, we might align this hypothesis

with the search for fundamental physical laws, such as laws about symmetry that explain and constrain relationships between fundamental subatomic particles. In the world of biology, we might align this hypothesis with the definition of a living entity by a genetic code that reproduces itself. Although life depends on a hospitable environment, its form depends on laws that are 'virtually' independent of environment.

Formism also gives us a conception of learning. Learning must be primarily an activity through which we uncover or deduce a system of universal relations and qualities. Consequently, it must involve abstract reasoning. A learning process that relies on abstract reasoning is no longer bound by concrete experience. Further, if we are willing to follow a theorist such as Arthur Jensen and derive a formist psychology from a formist biology, we might describe the capacity to reason abstractly as an innate, genetically determined phenomenon that is not dependent solely upon experience. The pattern of 'intelligence' as measured by an appropriate standardized test may reflect this 'fact.' Of course, one must remember that the test was designed to produce this pattern and that the pattern itself is an idealized form.

How does the world hypothesis of formism frame questions and answers about the assessment of science learning? First, we must understand a science curriculum as an attempt to reveal a reality that has form and laws about form that are accessible to human rational inquiry. For young children, this curriculum may simply be designed to build understanding of how to describe form and transformation in form through qualitative and quantitative concepts and relationships. For example, as noted earlier, as kids observe many tadpoles changing into frogs, they also can abstract the general form of a species. For older learners, a science curriculum may focus directly on the method of inquiry itself. Scientific experimentation is a result of an ongoing rational thought process that reflects the difficulty of abstracting an underlying reality of scientific law and form.

A formist may see the objective of a science curriculum as the construction of a learning environment in which children, either through logic or insight, uncover a stable reality of form and scientific laws that explain form. Developmentally, this curriculum should be built so that learners can deduce or infer qualitative and quantitative concepts as well as relationships between these concepts. In addition, a curriculum should give capable learners ample opportunity to develop a scientific way of thinking, one that relies upon logical abstract reasoning.

Mechanism

A second analytical hypothesis, the world hypothesis of mechanism, reduces the structure of all knowledge to the parts and operation of a machine. The metaphor that describes this hypothesis is the machine—in its full-blown form, a cosmic machine composed of fundamental parts. This cosmic machine must have principles of operation that define the working of its parts. The quest for knowledge is the quest for an understanding of this cosmic machine.

In a historical context, this world hypothesis relates to the philosophical work of empiricists such as David Hume and, in a more contemporary setting, the philosophy of Gilbert Ryle. A contemporary physicist might construe the world hypothesis of mechanism as supporting the search for a vast quantum electromagnetic gravitational field that, by virtue of the way it operates, is the cosmic machine. A contemporary biologist might similarly see mechanism as supporting the principles of molecular biology, according to which DNA, and the interaction between RNA and other biochemical agents, define the genetic makeup and reproduction of living entities. Place these “mechanical” living entities in an environment, and a larger mechanistic understanding is needed to describe an ecology in which species may live or die.

The world hypothesis of mechanism also frames a conception of learning. Learning must be a behavior of a person who is a mechanism, a behavior that involves the construction of knowledge from concrete experience. This definition of learning also points to a long history of mechanist psychology ranging from the physiological psychology of Wilhelm Wundt to the radical behaviorism of B.F. Skinner. The construction of a mechanist psychology that can explain how an individual who is part of the cosmic machine can come to understand that machine by learning has always been a difficult task.

How does the world hypothesis of mechanism frame questions and answers about the assessment of science learning? From a mechanist's viewpoint, a science curriculum reflects different levels of our understanding of the cosmic machine. A hierarchical representation of the sciences, beginning with physics, followed by chemistry, biology, and various social sciences, illustrates that our understanding is currently incomplete and fragmented. In addition, observation of scientists who work in these fields reveals a range of behaviors, including verbal behavior, that represents current knowledge and research questions in a field.

In this context, the major problems of science education involve understanding how a conceptual world of science knowledge and the process of scientific inquiry can be reduced to a set of human behaviors and behavioral rules. A scientist does not discover a reality independent of his or her own activity. For a mechanist, any complex conceptual system of scientific knowledge is a system of verbal behavior that generalizes concrete experiences that occur in the cosmic machine: Any process of scientific thinking or experimentation is reduced to behavioral action in the machine. A science curriculum must reflect the behaviors associated with someone who is 'doing science' and must create a learning environment in which children can reproduce and extend these behaviors. For example, as noted earlier, the objective of helping a child understand a general concept of metamorphosis may be connected with a science lesson that compels the child to observe several instances of frog development and to associate these observations with the word metamorphosis. In addition, disciplined observation must be taught within a larger repertoire of behaviors that are characteristic of scientific inquiry.

In general, a mechanist views the problems of science assessment as problems of understanding the types and sequence of behaviors that a science curriculum creates in the classroom. These behaviors are a consequence not only of an environment constructed by teachers but also of the unique behavioral history of learners. To be effective, a curriculum should bring all learners to a set of behavioral objectives.

Organicism

In contrast to the analytic world hypotheses of formism and mechanism, Pepper proposed two synthetic world hypotheses. Instead of beginning with parts and building the structure of knowledge, we begin with a whole or a process that must constrain the nature of its parts. The first of these world hypotheses that I will describe is organicism.

The metaphor that defines the structure of this world hypothesis is the living organism and its process of biological growth. A process of development and regulation resulting in a living organic whole is a metaphor for a process of knowledge that results in a structure of knowledge. An organicist defines the structure of all knowledge by looking at how it develops or by analyzing its history of development. In this context, a history of knowledge displays progressive stages, each of which resolves contradictions in a previous stage. Consequently, it also displays

movement toward more inclusive, ordered, and accurate states of knowledge. An organicist might go so far as to postulate a stage of knowledge where absolute truth has been, or will be, acquired.

Philosophically, this hypothesis can be associated with the work of G.W. Hegel and Herbert Spencer. The world hypothesis of organicism also frames a type of psychology and a conception of the process of learning. This is clearly illustrated in the work of Jean Piaget. Piaget defined a sequence of mental stages that emerge from a learning process embedded in an organic process of self-regulation. He called the learning process reflective abstraction, representing a psychological self-reflection and abstraction of physical and, eventually, mental activity. From this perspective, all knowledge must be framed by a process of human growth. In Piaget's terms, a philosophical conception of knowledge requires the construction of a genetic epistemology.

A physicist or biologist who takes an organicist perspective would emphasize that scientific knowledge is more than a state of knowledge about an independent reality. It is a stage of knowledge that has developed and will continue to develop through resolution of conflicting conceptions of scientific phenomena into more consistent and ordered models of reality. For example, a physicist might now say that an important problem in understanding the fundamental laws of physics is the integration of a description of gravitational force with nuclear and electromagnetic forces. A uniform conception of force would create a more consistent and ordered description of reality. However, this will undoubtedly lead to new conceptual confusions and contradictions whose resolution points to the continued growth of scientific knowledge. In general, the growth of scientific knowledge is toward a more coherent, consistent, and accurate description of reality.

In addition, a state of scientific knowledge and the process of its transformation reflect a process of human cognition. This is clear when we understand that the scientist's greatest capacity is to create a world of possibility and to test this world logically against a world of human experience. Learning by reflective abstraction and the emergence of hypothetical thinking mark the progression of scientific knowledge.

How does an organicist frame the problems of science education and assessment? An organicist first and foremost would understand a science curriculum as both an opportunity to broaden a child's knowledge about reality and an opportunity to expand the way a child thinks, since these processes are intrinsically linked. In fact, one might argue that a science curriculum should be placed primarily at the service of cognitive development. Some of the

fundamental concepts of science education, such as number, mass, volume, time, and conceptions of causality are intimately tied to mental development. In this context, science learning in a child's early years should involve challenging a child's conception of reality to promote mental development.

In an organicist's world, assessment of science learning should primarily be an evaluation of mental development. Does a science curriculum expand a child's current mental level in order to set the stage for transition to a higher level? Does a science curriculum create the disequilibrium that promotes transition to higher stages? In general, what we assess is how a curriculum facilitates changes in knowledge and in the thought process itself.

Contextualism

Contextualism is another synthetic world hypothesis. In this hypothesis, knowledge and a process of knowing are related metaphorically to the description of a historical event. The texture of such a historical event might be composed of a location or individuals in a location who act in some way. The quality of the event might be the inferred motivation or the reasons individuals give for acting the way they do. However, quality and texture emerge only when a person or historian places the event in a context.

Although context in a historian's work represents the need to re-present an event that occurred in the past to an audience in the present, in the world hypothesis of contextualism, it is a metaphor for any attempt to analyze or re-present reality. Contextualism has another important ingredient. Context, produced by deciphering a texture and quality of an event, is defined by a person. It is not part of an objective world existing independently of human activity. From a contextualist point of view, the structure of knowledge is fused with the intentions, purposes, and instrumental acts of human knowers.

The world hypothesis of contextualism can be related to the philosophical works of William James and John Dewey. In a different philosophical perspective, Karl Marx's conception of labor as an instrumental activity that defines society and individuals' conceptions of society is also an example of contextualism. In the domain of psychology, contextualism frames the works of the American psychologist James Mark Baldwin (who also used the term genetic epistemology), Lev Vygotsky, and, in a contemporary context, Klaus Riegel. For example, Riegel proposed a process of learning that depends on dialectical thinking. Although dialectical

thinking is a process of resolving contradictions in experience, this experience is both changed and comprehended for uniquely human purposes or intentions and, consequently, is dependent on a context. Learning is a problem-solving activity.

In the world hypothesis of contextualism, knowledge is both a description of a potential context and the instrumental activity that produces a context. Scientific knowledge is sometimes mistakenly characterized as a description of a reality that is independent of context; it is more accurately defined as a description of a potential reality that is context-dependent. One criterion of scientific knowledge, as highly criticized knowledge, is that it represents a conceptually abstract reality that can be translated into human experience—experience with roots in intentional or instrumental activity. This instrumental activity can change what is experienced—it can create new and novel contexts.

A history of scientific inquiry is a history of human instrumental activity. In a broad sense, this instrumental activity may be a natural adaptive activity or an activity that reflects the need for physical and social competence. This activity takes on many forms and is interwoven with many aspects of what a scientist does. For example, upon close analysis, we might find that a scientist's observations are defined by some type of measurement. Sometimes, the measurement activity is so interwoven in an event that a precise observation of some phenomena cannot be made. For example, in the subatomic world, a scientist can measure and define only probabilities of certain qualities and textures of events, such as the positions or amount of energy associated with high energy particle collisions. The scientific activity of observation can be understood as the activity of creating a context through measurement undertaken with some intent or goal.

Science is more than observation or measurement. Observations must enter into hypotheses that define concepts and predict the outcomes of further manipulations of experience. Scientific knowledge is a description of the texture and quality of a context that could be created by human activity. These hypotheses are not neutral constructions. They are designed within the collective human intention of describing and creating a necessary human context. A history of science is a record of a collective history of the broadening and refinement of contextual knowledge. It is also a history of methods of human inquiry that can become, in turn, the basis for new ways of thinking. It is a history of learning by problem solving.

How does a contextualist perceive the problems of science education and assessment? A science curriculum should re-present

significant science concepts or ideas by producing a context. To learn something new, children must be able to apply some problem-solving activity that can reveal the quality and texture of a context. For example, if a curriculum objective is for children to understand a concept of ecology, we may involve children in deciding what defines a livable environment for frogs or why certain changes in environment produce certain changes in development. These problems require the intentional construction of a context. On the one hand, the quality and texture of this context is a re-presentation of a collective history of science learning. On the other hand, the quality and texture of this context is a child's construction of a piece of science knowledge.

It is important to note that a contextualist, like an organicist, may talk about contradictions in experience that lead to, or facilitate, mental development. However, for a contextualist, the contradiction emerges from an instrumental activity and is resolved within a context. Consequently, scientific knowledge and an ability to think abstractly are not totally internal mental developments. Rather, their development is tied to the production of context, and contexts may be changed by one's activity. Thus, we should not be surprised to see a child think "abstractly" in one context and "concretely" in another.

WORLD HYPOTHESES AND SCIENCE ASSESSMENT

In Pepper's view, the world hypotheses of formism, mechanism, organicism, and contextualism delineate the grounds of philosophical debate about the structure of knowledge. I also believe they delineate several different positions about the assessment of science learning—positions that may not be resolvable by a better theory. In table 1, I summarize how each world hypothesis frames and answers questions about what is learned through scientific inquiry, how it is learned, and what should be assessed.

I hope the reader still feels that these potential conceptions of assessment all have a kernel of truth. The ideas of assessing a state of conceptual knowledge that describes a reality of form, a behavioral repertoire tied to concrete experience and representative of what a scientist does, a stage of knowledge and process of thinking that has a developmental history, and knowledge that evolves from the ability to produce and solve problems should all have commonsense appeal. I also hope the reader can feel the potential conflict between these positions. In some way, they are based on differing conceptions of what can be learned, how learning takes place and, more generally, of the person who can learn.

TABLE 1. Summary of Implications of World Hypotheses

<u>World hypothesis</u>	<u>What is learned?</u>	<u>How is it learned?</u>	<u>What should be assessed?</u>
Formism	A conceptual system that explains a reality of forms and laws about change in form.	Learning by abstract reasoning.	The present state of science knowledge of a child and the child's capacity for abstract thinking.
Mechanism	Conceptual knowledge that attempts to explain the parts and operation of a cosmic mechanism.	Learning by generalization from concrete experience.	A set of verbal and physical behaviors.
Organicism	An incomplete and contradictory conceptual system that is progressing toward greater coherence and accuracy.	Learning by reflective abstraction.	The structure of a child's thought as it moves toward abstract thought.
Contextualism	Conceptual knowledge that describes a reality that can be actualized within human experiences.	Learning by problem-solving.	The ability of a child to build or reproduce a context that defines and solves a scientific problem.

How can we account for these multiple, contradictory theories of assessment?

One answer involves understanding that the world hypotheses we have used to frame these different answers are limited in scope, internally inconsistent, and subject to refinement. Although we have not focused on the problems of interpretation within each world hypothesis, they do exist. For example, formism presents a problem about the meaning of existence. How can universal forms exist? In mechanism, how do we understand how we, as part of the cosmic mechanism, can have objective and shared knowledge of how it works? Contextualism, by definition, is open to change and novelty since the context of an event is a changing phenomenon tied to human activity. Finally, organicism necessitates suspending any doubt about a not-yet-achieved absolute state of knowledge, even though a current stage of knowledge has seemingly irresolvable contradictions. In this context, further refinement of one of these hypotheses or a hypothesis not yet proposed may save the day. Perhaps one hypothesis is more adequate than the others and will yield the answer.

Another answer may come from analysis of the origin and continued elaboration of world hypotheses in metaphor and in structural and multiplicative corroboration. First, we must assume that the four world hypotheses reflect different types of human relationships that frame human experience. The metaphors used to infer a structural relationship in knowledge are metaphors that capture different types of human relationships. Structural corroboration represents the continued interpretation and expansion of

these relationships in explanations of human experience. Consequently, the continued adequacy of four world hypotheses may indicate that four types of human relationships continue to be the basis of interpretation of reality. We might say that a person is simultaneously in relationships that reflect form, concrete experience, natural growth, and adaptation or mastery. I will call this possible explanation for the existence of different world hypotheses a theory of hermeneutic learning.

However, learning is not only an interpretation of relationships. It is also an interpretation that must be shared. Multiplicative corroboration clearly indicates that knowledge must be shared. Each world hypothesis represents a collectively agreed upon interpretation of human experience or is a product of intersubjective learning.

A learning process that is both hermeneutic and intersubjective may support the necessity of competing world hypotheses. A useful analogy is the process of dialogue. A person who learns engages in a dialogue with others. This dialogue requires an interpretation of human experience that can be shared and also allows particular types of interpretations. Since we cannot step out of dialogue to communicate, we are forced to explain experience within these types of interpretive frames.

Again, we are not in a position to pursue a lengthy discussion about world hypotheses. The point is that we may have grounds to consider multiple answers to a question about assessment of science learning. These answers cannot simply be pitted against each other because they reflect different, yet meaningful, conceptions of what is learned in a science classroom and how that learning takes place.

What are the implications of a theory of science assessment that proposes four different conceptions of assessment? I believe, first and foremost, that we are allowed a much more expansive way of looking at science education—a way that takes into account the complexity of learning that resists most theories of learning. Science learning is about understanding form and laws of change between form and this type of understanding requires abstract reasoning ability. Science learning is also about understanding a reality that behaves like a machine, and the primary evidence of the functioning of this machine is our own experience as part of the machine. In addition, science learning is intimately tied to a history of science that leads to a more accurate and coherent description of reality. It also is intimately tied to a process of mental development that leads to a form of hypothetical thinking. Finally, science learning is about solving uniquely human problems that emerge

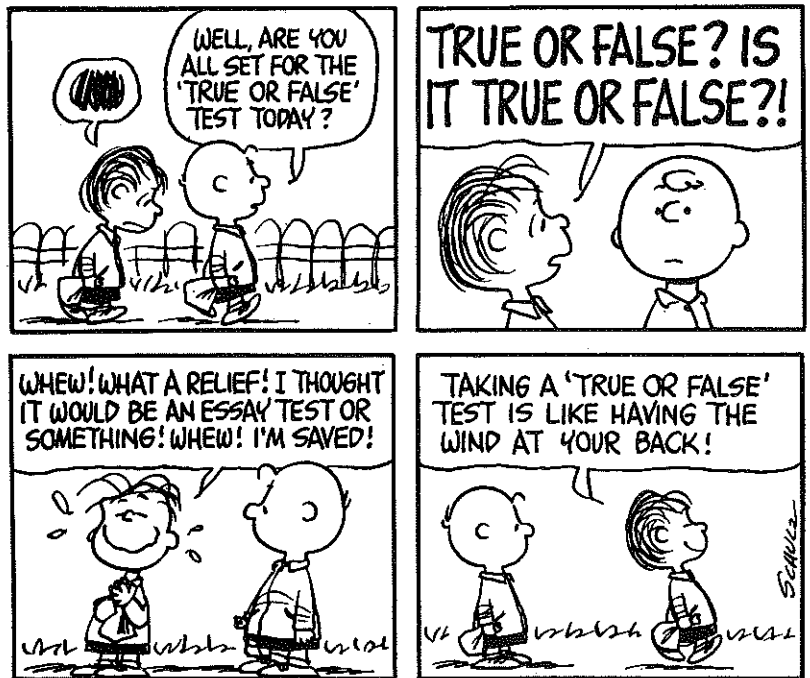
because of the need to master an environment. While this problem-solving activity changes the way we think, it also changes the problems we must think about.

A second implication of multiple theories of assessment is that we are forced to look at the limits of different types of frameworks. Pepper's theory of world hypotheses deals with theories that attempt to define the limits and structure of all knowledge. In this context, we expect theories that can expand beyond available evidence and explain the structure of all knowledge. However, we must be more careful when we look at the smaller implications of these theories. These implications must be considered in light of both continued hermeneutic and intersubjective development as well as alternative relational frameworks. For example, although we may be able to see the importance of modeling reality through the construction of forms and laws about forms, do we have evidence that potential differences in intelligence are hidden in a law about human forms and that these differences prefigure an ability to learn through experience? One idealized form, the normal curve, is often presumed to prefigure the ability to learn. Does this over-emphasize a formist construction of reality?

We may seem to be allowing ourselves to pick and choose between possible implications of theories framed by world hypotheses. However, remember that these theories require continued hermeneutic and intersubjective development. We must strive for interpretations of reality that have intersubjective validity and that further elaborate on human experience. In the context of assessing science learning, we must continue to refine our answers about what is learned, how learning takes place, and what should be assessed within the multiple dimensions of human relationships.

PART THREE

Large-Scale Assessments



Introduction

George E. Hein

Most large-scale assessments in the United States have relied exclusively on multiple-choice tests designed to optimize their psychometric parameters. The technical problems associated with large-scale assessments in particular favor this approach, since it appears to be most economical, least troubled by tester biases, easiest to administer, and most direct in providing comparison data with other populations and with past assessments.

Yet an increasing number of agencies have begun to explore alternatives to short-answer, paper-and-pencil tests for state and national assessments. The National Assessment for Educational Progress carried out experimental work in this area, and published a booklet, *Learning by Doing*, which describes the advantages of performance-based testing.

The two papers in this section address issues associated with developing alternative assessments on a large scale. Joan Boykoff Baron describes the Connecticut Educational Assessment Plan in Science, the state department of education's 1984-85 survey of students' science knowledge. This assessment used a combination of short-answer questions and other means. Baron's paper describes the general approach, some of the findings, and, most important for this volume, some of the striking differences found between types of assessment instruments.

The most ambitious large-scale science assessment in recent years to use a range of instruments, including considerable performance testing and open-ended paper-and-pencil questions, has been the work of the Assessment of Performance Unit (APU) in the United Kingdom. This group was charged with developing national surveys in all subject areas in the late 1970s and carried out national assessments for several years.

The APU testers interviewed children, asked them to complete actual science tasks, and gave them both short-answer and open-response written tests. The students' free responses to both the tasks and the open-ended questions gave the APU staff the challenge of struggling with the range of responses children produced. They discovered that students often grapple with problems different from those posed by the examiners, and that they bring to

bear experiences and preconceptions different from those imagined by the professionals who produced these carefully chosen assessment exercises.

Such problems are familiar to social scientists who follow more qualitative research traditions rather than relying on procedures based on statistical models. They repeatedly report that their work requires them to address a wide range of theoretical concerns. Issues related to the relationship between the instruments used and the respondents' approaches to these instruments (which may be quite different from what the researchers had intended) frequently play a large part in these discussions. Because qualitative methods involve researchers in intensive field work, where they deliberately interact with subjects in nonstructured ways, methods for dealing with unexpected outcomes are available.

In her paper, Patricia Murphy not only describes the APU approach and some of the results, but also discusses the complex issues that arise when assessment is more open, allowing respondents to explain their answers and help formulate the questions.

Both papers stress the importance of using a variety of assessment instruments in order to find out what children know and can do in science; both illustrate this point with data from actual large-scale assessments; and both demonstrate the possibility and value of having assessments based on more than multiple-choice, paper-and-pencil tests.

What We Learn from State Assessments of Elementary School Science¹

Joan Boykoff Baron

In 1984-85, The Connecticut State Department of Education conducted its second assessment of science in grades 4, 8, and 11. The science testing was part of the Connecticut Assessment of Educational Progress (CAEP) program. This chapter will describe both what we learned about the knowledge, skills, and understandings of elementary school students and what we learned about assessing elementary school science.

Ten years of experience in developing state assessment programs has convinced me that the ultimate criterion for any test is whether it truly represents what is most important for students to know and be able to do. Recently, statewide tests have taken on unprecedented importance and have become magnets for instruction. In states both with and without a statewide curriculum in science, many school-based educators believe that state tests embody the content and ideas that science educators think are important. This increased influence presents state test developers with both an opportunity to help guide school programs and a heavy responsibility to make the smallest number of compromises on the nature and the quality of the tests.

The following set of beliefs guided the development of the Connecticut Assessment of Educational Progress tests in science in 1984-85. To a large extent, they reflect what we had learned from our previous assessments in science and other areas:

- The assessment should include a hands-on performance component to assess the ability of students to conduct simple investigations, use measuring devices, and design experiments.
- The assessment should include some open-ended items that allow students to demonstrate their understanding of science concepts by generating responses.
- The assessment should include more than one type of item to assess skills and understandings in order to secure corroboration of the findings.

Joan Boykoff Baron, a former high school English teacher, is an education consultant at the Connecticut State Department of Education. Most recently, she has been developing high school science and mathematics assessments in which groups of students work together to solve complex, sustained problems. This paper was written while the author was a visiting scholar at the Center on Research Learning and Schooling at the University of Michigan, Ann Arbor.

- The assessment should include items that cover the full range of cognitive levels (e.g., knowledge, analysis, synthesis, etc.).
- The assessment should cover the scientific knowledge everyone needs to function effectively in a modern technological society as well as knowledge that people entering scientific or technological fields require.
- The assessment should cover a broad range of science knowledge and understanding, yet maintain a relatively small testing burden for individual students and schools. This could be accomplished through matrix sampling, which allows different students within a classroom to respond to different combinations of items.
- The assessment should include some items that had been administered nationally so that we could compare Connecticut students with students nationally.
- The assessment should include some items administered in previous Connecticut assessments so that we could monitor changes in students' performance.
- The assessment should administer some items to more than one age group to see the extent to which older students perform better than younger ones.
- The assessment data should be disaggregated by subgroups of interest (e.g., males and females, urban and suburban schools) so that we could monitor how any differences between these subgroups change over time.
- To augment the utility of the achievement results, student, teacher, and principal questionnaires should be administered concurrently with the assessments.
- Although only a sample of students statewide is included in the assessment, every superintendent should have the opportunity to participate in a local option to test some or all of the district's students.

The remainder of this paper provides support for these beliefs.

A DESCRIPTION OF THE SCIENCE ASSESSMENT

Each student in the statewide sample took a group-administered test composed of 49 multiple-choice and open-ended items. They and their teachers and principals also responded to a questionnaire. A subset of 300 of these students also participated in an individually administered practical test. The students participating in both components were selected from a stratified random sample of students throughout Connecticut. The Department contracted with Advanced Systems of Measurement and Evaluation, Inc., in Dover, New Hampshire, to select the sample, develop and score the tests, and analyze and report the results.

Group-Administered Tests

During the 1984-85 school year, two different test forms were administered at each grade with each student in the statewide sample completing one of these forms. As part of the statewide sample, we tested 2,696 fourth-grade students in 73 schools; 4,073 eighth-grade students in 57 schools; and 3,842 eleventh-grade students in 58 schools. Most of the test items were multiple choice; a few in each form were short answer, requiring hand scoring. Table 1 shows the distribution of test items across content categories for each of the three grade levels tested.

The content categories listed in table 1 was one of two dimensions we used to classify test items. Every item was also assigned to a cognitive process level—measuring knowledge, comprehension, application, or the higher process skills of analysis, synthesis, and evaluation. We were thus able to monitor whether the test questions as a set included a majority of items representing more than just recall of information. (I recognize that only the teacher knows the “true” cognitive level of a question, since any item could become a knowledge-level item if the material it covers was taught in the same format as the item).

Individually-Administered Tests

Practical exercises requiring the use of scientific apparatus were administered to a subset of 900 students in the statewide sample (i.e., 300 students at each grade level). These exercises, given to individual students, asked the students to conduct simple investigations, use measuring devices, and design experiments.

Questionnaires

Questionnaires were administered to all students, teachers, and principals in the statewide sample. They addressed matters

TABLE 1. Distribution of Items by Content Categories

Category	Number of Test Items		
	Grade 4	Grade 8	Grade 11
Scientific Inquiry			
Awareness of Science and Scientific Processes	3	4	9
Designing Experiments	3	5	5
Observing and Measuring	8	4	5
Interpreting and Translating Data	7	3	6
Drawing Conclusions and Inferences	4	5	4
Life Sciences			
Characteristics of Life	4	5	5
Animal Life	8	14	13
Plant Life	10	5	3
Ecology and Environment	4	8	9
Earth and Space Sciences			
Astronomy	7	7	7
Climate and Weather	5	5	7
Geology	6	7	9
Oceanography	2	5	3
Physical Sciences			
Physics	15	19	19
Chemistry	12	14	16
TOTAL NUMBER OF ITEMS	98	110	120

such as backgrounds, attitudes, and experiences; program and instructional characteristics; and materials and equipment used in the classroom. This information provided a richer context for interpreting the test results.

The Local Option

Every superintendent in Connecticut was offered the chance to participate in a local option that enabled districts to supplement the statewide sample with their own students. An additional 2,542 fourth-grade students in 49 schools; 2,920 eighth-grade students in 23 schools; and 3,096 eleventh-grade students in 23 schools were included in the testing through the local option

Test Administration

The bulk of the testing took place during two-week intervals in November, February, and April at grades 4, 8, and 11 respectively. The first week of each testing period was used for scheduled group testing, and the second week was reserved for make-up testing. The multiple-choice tests were administered to

intact classroom groups designated by our contractor. Local school personnel administered the group tests, following detailed instructions provided in an instruction manual. An additional subsample of 30 schools at each grade level participated in the practical testing of individual students. This part of the assessment was conducted after the group testing and was coordinated entirely by trained external test administrators who tested 10 randomly selected children at each school. The students were sent to a special training room two at a time. This practice had been used successfully by the Assessment Performance Unit in Great Britain² to reduce examinee anxiety. The two students worked independently on completely different exercises and were observed at the appropriate times by the administrator. After each pair of students finished, the administrator encoded their responses on machine-scorable answer sheets, using detailed scoring guides for the exercises.

WHAT WE LEARNED ABOUT OUR STUDENTS' KNOWLEDGE AND SKILLS: ACHIEVEMENT RESULTS BY SCIENCE DISCIPLINE

The CAEP assessment was a survey of statewide achievement in science. As such, its instruments were developed to cover as broad a range of concepts and topics as possible rather than a selected few "essentials." Although a few concepts were addressed by several items, this was typically not the case. The general character of such broad-gauged assessment instruments, then, makes the discussion of individual item results all the more important. We can have a great deal of faith in our ability to generalize from these results to how a larger population of students might have performed on these items because of the size and quality of the CAEP sample. However, when we attempt to generalize about the performance in subdomains of science on the basis of a few items, we must be more cautious. For that reason, average percentages correct for the different subtopics within science are not reported. We cannot claim that performance in one area of science is better than performance in another on the basis of such statistics. The items in one area might be more difficult. However, a close look at specific items and their patterns of incorrect responses, together with the subjective but informed interpretation of our Advisory Committee, has uncovered strengths and weaknesses of Connecticut students.

The major focus of the discussion throughout the following sections will be on the fourth-grade results—the only elementary

school grade tested in this assessment. However, where useful for placing the fourth-grade results in a broader context, the results from grades 8 and 11 are presented. Because Connecticut students are not substantially different from students nationally (see table 2), many of the findings presented here would probably be similar to those found in a large national sample. We will discuss the results for each separate category listed in table 1.

TABLE 2. Average Deviations in Percent Correct Between 1984 and 1985 Connecticut Results and Revison Connecticut, U.S., and Northeast Results

(n=number of items)

Topic	Grade 4			Grade 8			Grade 11		
	CT	US	NE	CT	US	NE	CT	US	NE
Scientific Inquiry	+3.5 n-6	+7.7 n-7	+4.7 n-7	+0.3 n-4	+5.8 n-4	+7.0 n-2	-5.0 n-4	-0.6 n-9	-4.0 n-9
Life Sciences	+2.0 n-4	+0.9 n-8	-0.1 n-8	-2.0 n-9	-0.3 n-11	-4.3 n-6	-0.1 n-8	0.0 n-10	-4.3 n-10
Earth and Space Sciences	+2.8 n-8	+5.1 n-7	+4.6 n-7	+0.8 n-4	+2.0 n-8	+0.3 n-6	-1.7 n-6	0.6 n-9	-1.2 n-9
Physical Sciences	+3.0 n-8	+1.0 n-11	-0.5 n-11	-2.0 n-10	+2.2 n-13	-1.5 n-10	+0.9 n-13	+2.1 n-10	+0.1 n-10
Total Test	+2.9 n-26	+3.3 n-33	+1.8 n-33	-1.3 n-27	+1.8 n-36	-1.0 n-24	-0.6 n-31	+0.3 n-38	-2.3 n-38

Scientific Inquiry

Several CAEP items related to scientific inquiry dealt with the awareness of science and science processes. A long-time concern of science educators has been the misconceptions and stereotypes students hold regarding science—beliefs that make science seem sterile and elitist and that may convince many students that science is not an area for them to pursue. The results of several items on the grade 4 test (figure 1) suggest that we still have reason for concern. Item 1-a suggests that many fourth graders associate science with conducting experiments and are less likely to see the roles of observation and measurement in science. Many fourth graders' answers to Item 1-b indicated that they believed scientists work only with hard facts and require laboratories.

FIGURE 1. Percent of Students Selecting Each Response to Science Process Questions

Item 1-a

Grades			
4	8	11	
17			Which of the following could a scientist do to find out how one rock is different from another?
14			A. observe
31			B. measure
38			C. experiment
1			* D. all of the above
			no response

Item 1-b

Grades			
4	8	11	
44			Which statement about scientists is TRUE?
26			* A. When scientists solve one problem, they often find new problems.
7			B. Scientists work only with facts, not with guesses.
24			C. Scientists know everything about what makes up things.
1			D. To do an experiment, a scientist needs a laboratory.
			no response

Two general conclusions can be drawn from the results on items about the design of experiments (figure 2). First, students show great faith in the written word in situations in which empirical verification is warranted. Item 2-a illustrates the point. Although two-fifths of the fourth graders (and two-thirds of the eleventh graders) recognized that measuring the height of all boys and girls in a school was the surest way to find out whether the boys are taller than the girls on average, large percentages of students at both grade levels opted for looking up average heights in an encyclopedia or in the school records showing the heights of students entering school. Item 2-b again shows students' tendencies to rely on nonempirical approaches. Forty-two percent of the fourth graders recognized that lifting heavier and heavier weights until a string broke is the best way to determine the strength of a string, but another 30 percent answered, "find out what the string is made of."

FIGURE 2. Percent of Students Selecting Each Response to Design of Experiments Questions

Item 2-a

Grades		
4	8	11
5		
15		
43		
21		
7		
8		
1		

In his school, Dan thinks the boys are taller, on the average, than the girls. Rose thinks the girls are taller. What would be the surest way to find out who is right?

- A. ask a teacher who knows all the students in school
- B. look up the average height of people of various ages in the encyclopedia
- C. measure the height of all the boys and girls in school
- D. look at the school records which tell how tall the students were when they entered school
- E. see who is taller, Dan or Rose
- F. I don't know.
- no response

Item 2-b

Grades		
4	8	11
9		
30		
6		
42		
12		

Which of the following is the BEST way to find out how strong a piece of string is?

- A. weigh the string
- B. find out what the string is made of
- C. take the string apart and soak the parts in water
- D. lift heavier and heavier weights with the string until it breaks
- E. I don't know.

On a set of items on observing and measuring, fourth-grade students had difficulty in reading a thermometer with multiunit gradations (figure 3).

Generally, on items measuring interpreting/translating data (figure 4), performance was higher when students were asked to read data than when they were asked to interpret it. Items 4-a and 4-b provide some interesting data on how well students can read the data from a bar graph. Item 4-b was difficult because it asked students for the fastest speed, requiring them to find the shortest time in seconds. Item 4-c provides some encouragement about students' ability to interpret data: Seventy-two percent of the fourth-grade students were able to select the correct interpretation of the data in the chart.

FIGURE 3. Percent of Students Selecting Each Response on Question about Observing and Measuring

Grades		
4	8	11
3		3
74		35
21		61
2		1

What temperature is shown on this thermometer?

- A. 17°
- B. 22°
- * C. 24°
- D. 27°

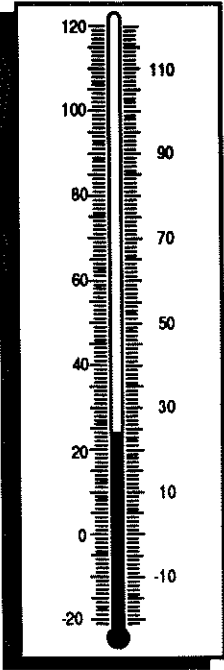
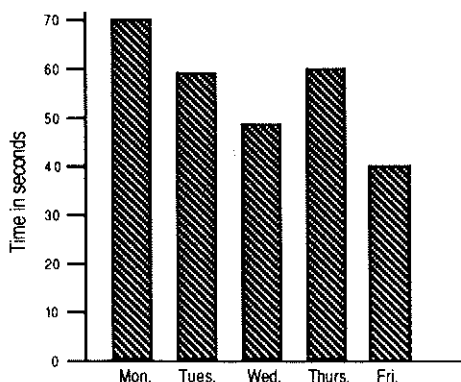


FIGURE 4. Percent of Students Selecting Each Response on Questions about Interpreting/Translating Data

Item 4-a

Blast O'Wind is a race horse that runs around a track each day. The graph below shows the time it took Blast O'Wind to run around the track each day.



How many seconds did it take Blast O'Wind to run around the track on Tuesday?

Grades		
4	8	11
3		
6		
86		
5		
1		

- A. 40 seconds
- B. 50 seconds
- * C. 60 seconds
- D. 70 seconds
- no response

Item 4-b

4	8	11
54		
3		
3		
39		
1		

Look at the graph again. On which day did Blast O'Wind run the fastest?

- A. Monday
- B. Wednesday
- C. Thursday
- * D. Friday
- no response

Item 4-c

A doctor kept records of breathing rates of people when they were resting. He made the chart below.

BREATHING RATES	
Person	Breaths in a minute
Baby boys	38
7-year-old girls	25
7-year-old boys	25
10-year-old boys	20
Mothers	16

The chart suggests that:

Grades		
4	8	11
8		
4		
8		
72		
7		
1		

- A. boys breathe faster than girls.
- B. girls breathe faster than boys.
- * C. older people breathe faster than younger people.
- D. younger people breathe faster than older people.
- E. I don't know.
- no response

Performance on items requiring students to draw conclusions and inferences (figure 5) was somewhat mixed. Sixty-nine percent of the grade 4 students recognized that the best they could conclude from the problem described in Item 5-a is that "some pieces of wood sink in water." However, when presented with a free-response item asking them to generate the conclusion they could reach from the data, students were much less successful. On Item 5-b, only 13 percent of the fourth graders wrote a valid conclusion (e.g., the heart beats faster when running). Seventeen percent provided an unwarranted conclusion (e.g., walking is best), and another 16 percent simply described what the graph is about, but did not draw a conclusion. This open-ended exercise provides strong evidence that students need more practice at drawing their own conclusions rather than selecting conclusions from a set of alternatives.

FIGURE 5. Percent of Students Selecting Each Response to Questions about Conclusions and Inferences

Item 5-a

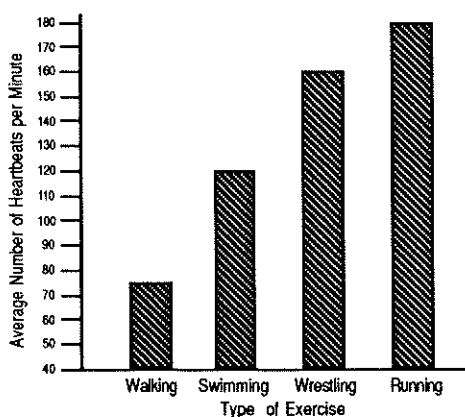
Bill had a piece of wood from a certain tree. When he put it in water, it sank. Which one of these did he find out?

Grades		
4	8	11
12		
4		
69		
6		
10		

- A. All pieces of wood sink in water.
- B. All pieces of wood float in water.
- C. Some pieces of wood sink in water.
- D. Some pieces of wood float in water.
- E. I don't know.

Item 5-b

Look at the graph below. What can you learn from it about the different types of exercise?



Grades			
4	8	11	
13			* A= valid conclusion (e.g., heart beats faster when running or causes heart beat to speed up)
17			B= unwarranted conclusion (e.g., walking is best, most or least healthy...)
7			C= translation (simply reading graph--heart beats 73 times per minute when walking, etc.)
0			D= combination B & C (e.g., 70% like to walk)
4			E= incorrect translation (e.g., running requires most breaths)
16			F= says what graph is about, but doesn't make conclusions or translation.
31			G= other incorrect response
12			no response

Life Sciences

The items measuring characteristics of life indicate that grade 4 students have only a fair understanding of the differences between living and nonliving things. Although 68 percent of the students recognize that a wooden chair is made from material that was once alive, the remaining 32 percent indicated that a metal cabinet, a stone wall, or a glass window was once alive (figure 6).

FIGURE 6. Percent of Students Selecting Each Response to Questions about Characteristics of Life

Grades			
4	8	11	
68			Which object is made from material that was once alive?
10			* A. a wooden chair
14			B. a metal cabinet
7			C. a stone wall
1			D. a glass window
			no response

On items about animal life, most fourth graders (93 percent) knew that a caterpillar (shown in a picture) would grow up to look like a butterfly (also shown in a picture). However, only two-fifths of these students knew where a baby chick growing inside an egg gets its food (figure 7).

When asked about nutrition (figure 8), only 47 percent of the fourth graders knew that fruits were important because they contain essential vitamins. Almost an equal number indicated that the foods pictured were protein foods, used for building muscle (see item 8-a). Fifty-three percent of the fourth graders could identify the most balanced menu from a set of four menus (see item 8-b).

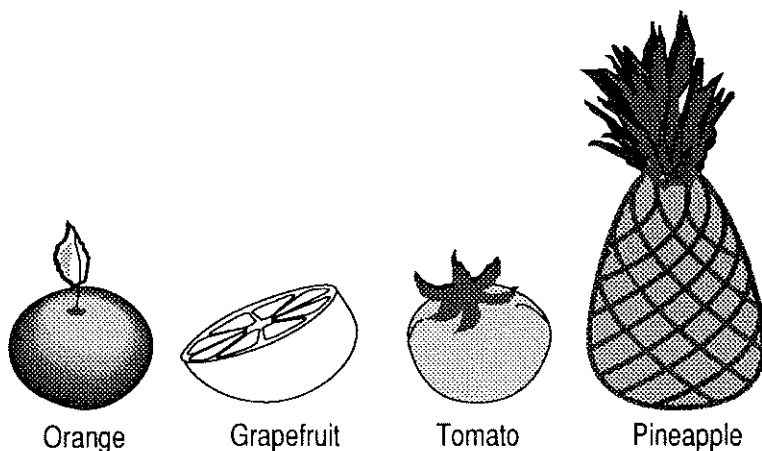
FIGURE 7. Percent of Students Selecting Each Response to Questions about Animal Life

Grades			
4	8	11	
17			A baby chick grows inside an egg for 21 days before it hatches.
15			Where does the baby chick get its food before it hatches?
11			A. It is fed by the mother hen.
41			B. It doesn't need any food.
16			C. It makes its own food.
			* D. The food is stored in the egg.
			E. I don't know.

FIGURE 8. Percent of Students Selecting Each Response to Nutrition Questions

Item 8-a

Look at the picture below:



Grades			
4	8	11	
3			The MOST important reason for having some of these kinds of foods daily is:
1			A. they taste refreshing.
47			B. they help you gain weight.
2			* C. they contain an essential vitamin.
43			D. they help you wake up in the morning.
4			E. they are protein foods, which build muscle.
			F. I don't know.

Item 8-b

Which one of the following menus is the most balanced?

A.

Menu A

Cereal
Fried potatoes
Bread and butter
Rice pudding
Iced tea

B.

Menu B

Green salad
Carrots
Meat loaf
An apple
Milk

C.

Menu C

An egg
Bacon
Sausage
Sliced cheese
Milk

D.

Menu D

Sliced tomatoes
Carrots
Baked potato
Cherry pie
Orange juice

Grades		
4	8	11
6		
53		
29		
7		
5		
1		

- A.
- B.
- C.
- D.
- E. I don't know.
no response

Test items covering emergency first aid procedures suggest that some myths need dispelling. Only 43 percent of the fourth-grade students recognized that covering a wound with a cloth and pressing firmly is the appropriate procedure for dealing with a rapidly bleeding cut. As figure 9 reveals, 48 percent of the students felt that the first thing that should be done is to wash the cut carefully. In this instance, the emphasis on the importance of cleanliness would result in inappropriate behavior.

FIGURE 9. Percent of Students Selecting Each Response to First-Aid Questions

Grades		
4	8	11
48		
3		
43		
2		
4		

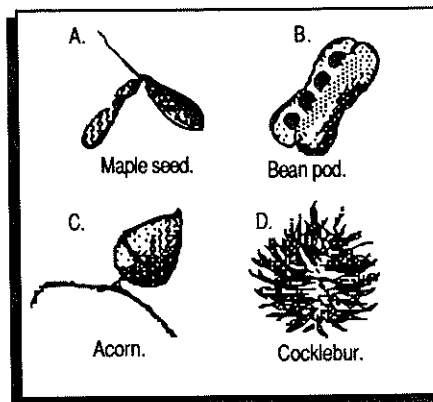
You and a friend are playing in a park. Your friend falls down and cuts herself on a piece of glass. The cut is bleeding a lot. What should you do first?

- A. get some water and wash the cut out carefully.
- B. have your friend hold her breath because a cut will not bleed if you are not breathing.
- C. put a clean cloth over the cut and press down firmly.
- D. have your friend run in place so her heart will beat faster and keep the blood inside her body.
- E. I don't know.

On items measuring knowledge of plant life, 67 percent of the students identified the picture of the cocklebur in figure 10 as the seed that would be spread by sticking to the fur of animals. At all three grade levels, performance was poor on items dealing with parts of the plant and photosynthesis (figure 11). At grade 4, 89 percent of the students identified the roots as the part of the plant holding the plant in the soil, but only 44 percent knew that seeds were produced in the flowers, only 40 percent knew that the leaves take in sunlight and produce food, and only 39 percent knew what is inside a seed.

FIGURE 10. Percent of Students Selecting Each Response to Question about Plant Life

Sometimes seeds stick to animals and are carried to new places where they will later grow. Which of these seeds would most likely be spread this way?



Grades			
4	8	11	
12			A.
7			B.
7			C.
67			D.
7			E. I don't know.

FIGURE 11. Percent of Students Selecting Each Response to Questions about Plant Life

Item 11-a

Grades			
4	8	11	
15			What part of a plant produces seeds?
11			A. roots
18			B. leaves
4 4			C. stems
11			D. flowers
			E. I don't know.

Item 11-b

Grades			
4	8	11	
15			What part of a plant takes in sunlight and makes food for the plant?
4 0			A. roots
22			B. leaves
13			C. petals
9			D. flowers
			E. I don't know.

Item 11-c

Grades			
4	8	11	
39			What is found inside a seed?
36			A. a young plant and stored food
9			B. many smaller seeds
15			C. a flower and little leaves
1			D. nothing--seeds are hollow
			E. I don't know.

Just over half the students knew that a cotton shirt originally came from a plant or that green plants are important to animals because the plants provide the animals with food and oxygen. Finally, students answered an open-ended item, "What things do plants need in order to produce food? (Name as many as you can.)" Most fourth graders included water (82 percent) and light (65 percent). However, only 27 percent included air, and 1 percent listed chlorophyll.

Performance on items assessing ecology and the environment (figure 12) indicates that students have only a superficial

knowledge of these topics. Only 47 percent of the fourth graders recognized that depletion of nutrients in the soil was a problem when plants grew less well year after year in a particular garden (see Item 12-a). On an open-ended exercise, only 19 percent of the fourth graders and 50 percent of the eighth graders were able to provide an example of a food chain composed of at least three links. A multiple-choice item asked students to indicate which of the following would cause a river to become polluted—fertilizer from farms, soapy laundry water, and chemical wastes from factories. Thirty-five percent of the fourth graders, 61 percent of the eighth graders, and 71 percent of the eleventh graders correctly answered “all of the above.” As shown in Item 12-b, the most popular choice among fourth-graders was “chemical wastes from factories.”

FIGURE 12. Percent of Students Selecting Each Response to Ecology Questions

Item 12-a

Grades		
4	8	11
15		
10		
47		
11		
15		
1		

- John's family has a garden each year in the same place. Each year the plants do not grow as well as the year before. This is most likely because
- A. the air is changing.
 - B. the Sun shines less each year.
 - C. some things in the soil that the plants need are being used up.
 - D. the plants have not been given enough water to meet their needs.
 - E. I don't know.
- no response

Item 12-b

Grades		
4	8	11
5	2	
8	4	
50	33	
35	61	
2	1	

- Which of the following would cause a river to become polluted?
- A. a farmer spreading fertilizer over a field next to a river
 - B. a house's soapy laundry water and other dirty water running into a stream that flows into the river
 - C. a factory dumping chemical wastes into the river
 - D. all of the above
- no response

Earth and Space Sciences

Our assessment results suggest that knowledge of specific facts is covered well in instruction in astronomy. Over four-fifths of the students at grade 4 knew that day and night occur because

the earth turns. Just under two-thirds of these students knew that the sun was the largest body, and just over two-thirds knew that the sun is a star. Approximately four-fifths of the fourth graders knew that the earth moves around the sun. Questions requiring higher thought processes yielded considerably lower results. Asked why telescopes are often built on mountain tops, only one-fifth of the fourth graders responded correctly, "to get above smog, dust, and fog." Three-fifths of the students selected as their reason that they would be "as close as possible to the stars," indicating a weak concept of relative distances in space.

The results on questions about climate and weather indicate that students have knowledge of many basic facts. Over half of the grade 4 students knew that frozen rain is sleet; 83 percent knew that the most damaging common characteristic of hurricanes and tornadoes is high winds.

On the geology/oceanography items, results were mixed. Fifty-one percent of the fourth graders believed that gasoline is the most plentiful fuel in the United States. Only 18 percent indicated that coal was most plentiful. Over half knew that as a diver goes farther below the ocean's surface, the water gets colder and the pressure becomes greater.

Physical Sciences

Several of the chemistry items dealt with states of matter. Sixty-four percent of the grade 4 students knew that water changing from a liquid to a gas is evaporation. Forty-four percent of the grade 4 students knew that water drops appearing on the outside of a cold glass of lemonade came from water in the air. On a set of items covering chemical changes, 45 percent of the fourth graders could identify "a candle burning" as a change in which a different substance is formed. Sixty-nine percent knew that iron would be more likely to rust when it is damp, and 68 percent knew that blowing on a campfire could speed the burning.

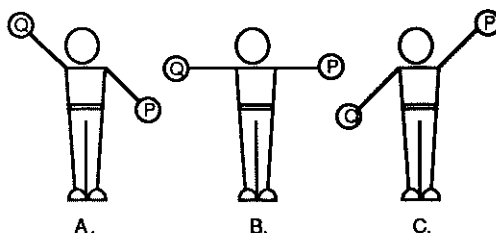
On the physics items, 28 percent of the grade 4 students knew that sand is used on icy roads to create friction between car wheels and the road. Half of the students claimed that the purpose of the sand is to help melt the ice. Many students may be confusing sand with salt. Ninety-one percent of the grade 4 students knew that a magnet would pick up an iron nail as opposed to a glass marble, a piece of paper, or a plastic comb; only 37 percent knew that magnetism rather than heat, gravity, or air pressure is what allows us to use a compass.

DIFFERENT PATTERNS OF RESULTS FROM MULTIPLE-CHOICE AND PERFORMANCE ITEMS

The results on the physics items include two interesting paradoxes. The first is portrayed in figure 13, which provides some of the most interesting data on the assessment. Seventy-one percent of the fourth graders answered this question correctly as compared with 56 and 57 percent of the older students. Have we succeeded in teaching our elementary school students a concept that baffles our older students? Unfortunately not. On the practical test, when asked whether a penny or a quarter would fall faster when dropped, only 5 percent of the fourth graders indicated that both would fall at the same rate. This is quite inconsistent with the results on the multiple-choice question.

FIGURE 13. Percent of Students Selecting Each Response to a Physics Question

Suppose that you want to drop a penny and a quarter at exactly the same time and have them hit the floor at exactly the same time. Which picture BEST shows how you would hold the penny and the quarter just before you drop them.



Grades			
4	8	11	
17	31	30	A.
71	56	57	B.
8	8	9	C.
4	3	3	D. I don't know
1	1		no response.

We can offer any number of interpretations of this inconsistency. Perhaps on the multiple-choice item, the fourth graders responded with what they had been taught, not what they really believed. When presented with the question prior to testing out their answer on the practical test, they may have expressed their

true belief. Another possible explanation of the multiple-choice item result is that the fourth graders read less into the question and chose the figure with the penny and quarter held at the same height because it “seems right” if you don’t realize that the penny and quarter weigh different amounts. The older students, however, might have considered the weight factor, which might have lead them astray. In conducting the investigation to determine what would actually happen, 94 percent of the older students correctly controlled the height of the two coins when released.

In the item reproduced in figure 14, students were asked to identify the circuit that would result in a bulb’s lighting when the switch was closed. Only 46 percent could do so. Interestingly, the practical test results show that 85 percent of the fourth graders could make a bulb light when presented with a battery pack with wire leads, a light bulb with wire leads, and insulated wires with alligator clips for connectors and time to try different configurations. This combination of items indicates that students cannot recognize a picture of a simple closed circuit, but they can “figure it out” (using trial and error) if provided with the opportunity. (If we were to use this practical test exercise again, it would be interesting to observe how many different configurations the children tried before completing the circuit.) This pair of items provides additional support for the importance of measuring scientific knowledge and skills with multiple approaches. Furthermore, it emphasizes the importance of being clear about what one is interested in knowing when constructing a performance item. Our exercise confounded knowing how to complete a circuit with being able to figure it out when provided with multiple opportunities.

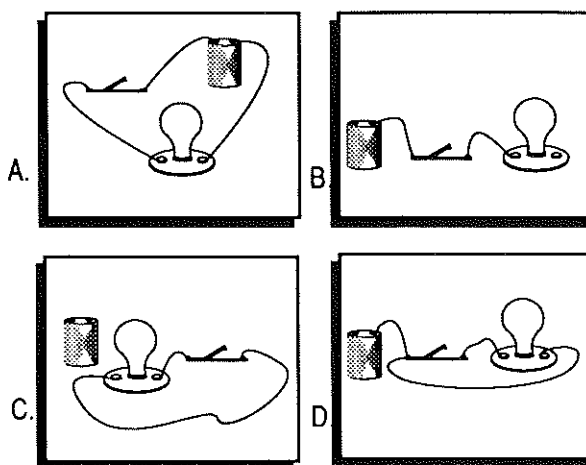
THE PRACTICAL TEST: SOME ADDITIONAL FINDINGS

Fourth-grade students had no difficulty sorting and classifying a variety of small objects (e.g., nuts, bolts, and washers that were either large or small and made of either brass or steel) into different groups. The fourth graders also did well on some measurement tasks. Asked to measure the length of a wooden block to the nearest centimeter, 91 percent of the students were successful. Asked to read the temperature of a container of cold water and a container of hot water, 85 percent of the students gave measurements within two degrees of the administrator’s readings. (The thermometers were marked in one-degree gradations, in contrast to the multiple-choice item shown in figure 3, in which the thermometer was marked in two-degree gradations.) Another question required students to determine the difference between the two water temperatures. Eighty-six percent of the students correctly

computed differences based on their readings. A follow-up question asked students to guess what the temperature would be if they mixed a larger quantity of cold water with a lesser quantity of warm water. One-half of the students correctly guessed a temperature between the warm and cold water temperatures and closer to the cold water temperature. Another 24 percent guessed temperatures in between the two extremes, but not closer to the cold water temperature.

FIGURE 14. Percent of Students Selecting Each Response to Question About Circuits

Look at the pictures below. Each shows a battery, a bulb and a switch. Which bulb will light when the switch is closed?

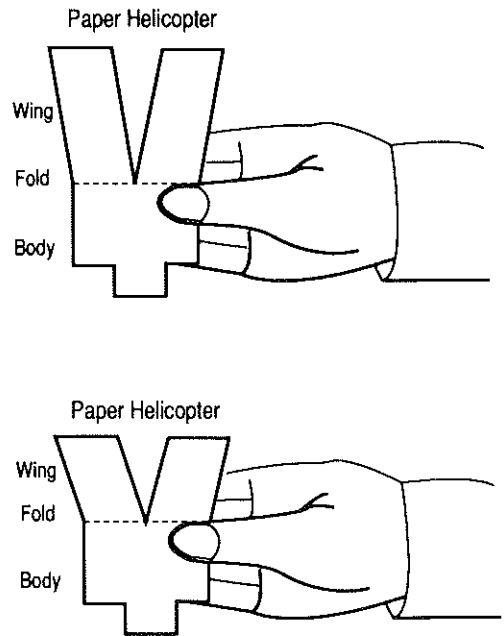


Grades		
4	8	11
46	65	75
21	13	11
5	4	1
16	13	8
11	5	5
1		

- * A.
B.
C.
D.
E. I don't know.
No response.

The students were also given a pair of paper "helicopters" (figure 15). They were asked to determine if anything was different about the way the two helicopters fell. Seventy-eight percent correctly noted that they fell at different speeds or with different rates of spin. As a follow-up question, the administrators asked the students why they fell differently; 73 percent of the students correctly attributed the different motions to wing length—the only difference between the two helicopters.

FIGURE 15. Paper Helicopters Given to Students in the Practical Test



Another simple investigation provided the students with a timer and an apparatus enabling them to set up pendulums with bobs of different weights and varying string lengths. Asked to time 40 complete swings of various pendulums, 89 percent of the students timed them reasonably accurately—i.e., they started and stopped the timer and counted swings appropriately. The students were next asked what they could conclude about pendulums. Sixty-five percent correctly concluded that shorter pendulums swing faster. Eight percent concluded the opposite, and 24 percent gave conclusions not comparing speeds.

COMPARISONS TO PREVIOUS STATEWIDE, NATIONAL, AND REGIONAL RESULTS

Although it is tempting to believe that direct comparisons to national norms can be obtained by administering sets of items common to statewide and national assessments, procedures are so different that any results must be viewed with extreme caution. First, most of the science items for which data were available were administered nationally at a considerably earlier period (1976-1977) by the National Assessment of Educational Progress

(NAEP) program, although a few items date from 1981-82. Second, NAEP sends teams of trained administrators to the schools where the testing is to be done, whereas teachers in CAEP generally administer the multiple-choice tests to their own students. The comparisons reported in table 2 are based on the most comparable data that can be obtained statistically. In general, differences of fewer than four percentage points should be considered inconsequential.

The most striking feature in the data is the great consistency among previous national, regional, and statewide results. At grade 4, Connecticut students show signs of strength relative to their national and regional counterparts in the areas of scientific inquiry and earth/space sciences. As for change over time, only the grade 4 students showed a small, but consistent, improvement since the previous assessment of science in 1979-80. On the average, grade 8 and grade 11 students changed very little.

Performance by Gender and Type of Community

Table 3 shows the average science test scores for males and females as well as for students from various types of communities. Again, certain cautions should be kept in mind when interpreting these findings. First, the reader should remember that small differences between achievement groups should not be overinterpreted. Second, observed differences do not indicate cause-effect relationships. Although a relationship between achievement and a given variable (e.g., gender of student) may exist, the data do not indicate the reason or cause for the relationship.

TABLE 3. Achievement Results by Science Area by Gender and by Type of Community

Science Area	Average Percentage of Questions Answered Correctly					
	Grade 4		Grade 8		Grade 11	
	Male	Female	Male	Female	Male	Female
Scientific Inquiry	56	55	56	58	53	55
Life Sciences	56	54	52	50	52	52
Earth/Space Sciences	52	48	47	43	52	45
Physical Sciences	56	51	54	47	54	46
	55	52	52	49	53	49
Total Science	55	52	52	49	53	49
Type of Community	Grade 4		Grade 8		Grade 11	
Large Cities	42		41		38	
Fringe Cities	58		51		52	
Medium Cities	51		50		51	
Small Towns	58		55		55	
Statewide Average	53		51		51	

On the average total test scores, males outperformed females by between 3 and 4 percentage points at all three grade levels. However, when looked at by science area, a more complex pattern emerges. For example, in scientific inquiry and life sciences, male and female performance levels were quite similar. Only in earth/space sciences and physical sciences did males have a sizable edge, particularly at grade 11. The grade 11 differences might be explained in part by differences in the course-taking patterns of males and females: A greater percentage of males than females had enrolled in general science, earth science, chemistry, and physics. Only in biology had more females than males been enrolled. At grade 8, where the differences in earth/space sciences and physical sciences are evident though somewhat smaller, course-taking patterns are a less tenable explanation.

Students in "big cities" (populations of more than 100,000) scored well below statewide averages at all three grade levels. Students in "small towns" (populations less than 25,000) scored somewhat higher than students statewide, and students in "fringe cities" (towns contiguous with large cities and with populations of over 10,000) and "medium cities" (populations between 25,000 and 100,000 and not fringe cities) generally scored very near the statewide averages. The exception to the latter was the grade 4 students in fringe cities who performed approximately the same as the grade 4 students in small towns. For the most part, the patterns of findings for type of community in table 3 are consistent with the results of other Connecticut testing programs.

Student Questionnaires

Student questionnaires permitted us to collect data on students' attitudes toward science. These attitudes become less positive as students progress through the grades. Seventy-four percent of the fourth graders, 57 percent of the eighth graders, and 40 percent of the eleventh graders reported that science was their favorite or one of their favorite subjects. Approximately three-fourths of the fourth graders believed that science will be very useful to them when they grow up. Sixty-nine percent of the eighth graders and 57 percent of the eleventh graders either strongly agreed or agreed with a statement about science being useful to them in future work. To help put these results in perspective, note that students report liking all subjects less as they proceed through school. Similar patterns were found in other content areas we have assessed since 1982 (e.g., social studies and English language arts).

Teacher Questionnaires

Teacher questionnaires provided information about teachers' science backgrounds, the amount of time they spend teaching science, their questioning practices, and their priorities for staff development. Although almost all grade 4 science teachers have had at least one college course in biology, a third to a half of these teachers have completed no college coursework in chemistry and physics. Five percent of the fourth-grade teachers had an undergraduate major in science. This number increased to 52 percent in the eight grade, and to 77 percent for teachers in grades 9-11. Over a third of the grade 4 teachers have not participated in workshops in science or in science teaching in the past seven years.

Grade 4 teachers were asked how many minutes per week their students receive science instruction. In many schools, relatively little time is spent on science instruction at grade 4. A third of the teachers spend less than 80 minutes per week and another third spend between 81 minutes and 2 hours per week (see table 4).

TABLE 4. Amount of Instruction per Week in Fourth-Grade Classes

Amount of Science Instruction per Week	Percent of Grade 4 Teachers Selecting Response
40 minutes or less	9
41 - 80 minutes	27
81 - 120 minutes	35
121 - 160 minutes	17
more than 160 minutes	7

One set of questions asked about the types of questions teachers at all three levels ask aloud in class and on written tests. Teachers were asked to estimate the proportion of questions of different types: questions requiring students to repeat memorized facts, questions requiring students to explain something, and questions requiring students to apply their knowledge to arrive at an answer. The students at grades 8 and 11 were also asked about their teachers' questions (see table 5). The general pattern is for greater emphasis on lower-level questions, both in class discussions and on written tests. The grade 8 and 11 students estimated greater frequency of explanation and application questions than did their teachers.

TABLE 5. Types of Questions Teachers Ask

Questionnaire Items	Grade 4			Grade 8			Grade 11		
	Percentage of Teachers	Ave. Test Score	Percentage Stud.	Tchrs.	Ave. Test Score	Percentage Stud.	Tchrs.		
OUT LOUD DURING CLASS:									
How many of your questions require the students to give back only facts and other information they have memorized?									
almost all	6	48	14	3	46	23	6		
more than half	26	51	21	23	50	26	19		
about half	38	51	40	44	50	37	36		
less than half	20	54	18	21	56	17	28		
almost none	3	48	6	7	53	7	10		
How many of your questions require the students to explain something?									
almost all	6	50	16	10	50	12	8		
more than half	22	53	29	23	51	26	26		
about half	38	51	34	36	51	38	35		
less than half	27	50	17	29	54	20	27		
almost none	1	39	4	1	42	5	1		
How many of your questions require the students to apply many things that they know in order to arrive at an answer or solve a problem about something they did not know before?									
almost all	3	52	9	5	53	11	6		
more than half	14	51	20	16	53	21	18		
about half	24	49	31	24	49	26	22		
less than half	45	53	29	48	51	29	43		
almost none	8	49	12	6	49	12	9		
ON WRITTEN TESTS:									
How many of your questions require the students to give back only facts and other information they have memorized?									
almost all	14	51	22	10	51	19	10		
more than half	33	52	29	31	51	27	24		
about half	31	50	30	37	49	30	30		
less than half	12	49	15	17	54	17	26		
How many of your questions require the students to explain something?									
almost all	1	47	11	2	45	8	3		
more than half	13	50	24	10	49	20	16		
about half	26	52	32	35	49	34	30		
less than half	40	52	26	42	56	27	42		
almost none	13	51	8	10	54	10	8		
How many of your questions require the students to apply many things that they know in order to arrive at an answer or solve a problem about something they did not know before?									
almost all	3	50	9	2	55	13	6		
more than half	9	49	17	9	52	18	16		
about half	14	49	28	20	48	28	17		
less than half	43	52	29	50	52	25	42		
almost none	25	53	17	16	51	16	17		

The last question on the teacher questionnaire at each level was a free-response question asking the teachers to list the two or three areas in which they felt inservice training would be most beneficial. Although many topics were mentioned, the agreement across grade levels on those most often mentioned was substantial. Most teachers want ideas—ideas for laboratory activities and for teaching various topics or concepts. At all grade levels, particularly at the upper levels, teachers want to learn more about the use of computers, both as a tool for science and as an instructional aid. Many teachers, given the opportunity to write their opinions, provided unsolicited views. Generally, the teachers were expressing their concerns about two matters—time to prepare for science instruction and money (i.e., salaries and funds for supplies). Some commented on the ineffectiveness of inservice-training activities and proposed a variety of means by which teachers could exchange ideas.

Principal Questionnaires

Principal questionnaires provided additional information about the amount of time teachers spend teaching science and whether this has changed over the last five years. The information provided by elementary school principals on the number of minutes per week students receive science instruction was consistent with the information provided by the teachers. When asked how often science instruction is provided to fourth-grade students, 15 percent of the principals indicated “every day,” 55 percent indicated “3 or 4 times per week,” 18 percent indicated “once or twice every week”, and 11 percent indicated “in blocks alternating with other subjects”. Ninety-five percent of the junior high principals indicated that their eighth graders have five class periods each week in science.

When asked to contrast the emphasis on science now as compared to five years ago, 56 percent of the elementary school principals reported a recent increase, as did 49 percent of the junior high principals and 68 percent of the high school principals. Almost all the remaining principals indicated that the emphasis has remained constant. Asked if their schools have a teacher who is given released time or additional salary to purchase equipment or help other teachers in science, only 19 percent of the elementary principals said yes, compared to 49 percent at the junior high level and 71 percent at the senior high level.

WHAT WE HAVE LEARNED ABOUT ASSESSING ELEMENTARY SCHOOL SCIENCE FROM ASSESSING WRITING

In this section of the paper, I will reflect upon experiences in another curriculum area—writing—in order to derive principles that are important in designing future assessments of elementary school science.

Just 15 years ago, state departments of education and local school districts were skeptical about using writing samples to assess students' writing. Instead, they used reliable, efficient multiple-choice tests to test the editing skills of students. Today, at least 18 states and thousands of local school districts use writing samples to assess their students' writing skills. What caused this change? Perhaps the largest single determinant was that training techniques were developed to score students' writing reliably. Procedures were developed to enable raters to apply a set of scoring criteria consistently. First, groups of experts selected anchor papers for each score point on the scale. These anchors, or range finders, served as models to train the raters. Raters practiced by reading students' papers and determining which range-finder paper most closely resembled the paper under consideration. Raters examined models of students' papers at all points along the scoring scale. Whether using holistic scoring (which generally focuses on an overall impression of both content and mechanics), primary trait scoring (which generally focuses on one major aspect of content), or analytic scoring, (which generally focuses on a set of separate dimensions of content and mechanics), trained raters were able to recognize a 1 paper as distinct from a 2, 3, or 4 paper. This resulted because all raters were looking for the same traits when they read the students' papers. In this way, two or more raters assigned the same rating to the same paper, a psychometric trait referred to as interrater reliability. This instilled a sense of confidence in educators and the public that writing could be scored objectively and efficiently.³

Once the reliability and feasibility issues had been settled, a ground swell of support arose for using writing samples to assess writing. This support was founded on issues of validity. The first of these is ecological validity. More specifically, teachers believed their students needed to develop writing skills because they would need those skills both in their further schooling and in the world outside of school. Therefore, the writing assessment should closely resemble the writing required in these settings—a criterion not met by multiple-choice tests.

Secondly, there were issues of content and instructional validity. Teachers believed in the importance of the writing process

and wanted to encourage their students to write as often as possible. Therefore, teachers needed accountability measures that would both preserve the direct teaching of writing in their classrooms and be sensitive to the growth of their students. Teachers believed that multiple-choice tests were inappropriate for assessing writing. Multiple-choice tests, they argued, measured different skills (e.g., recognizing grammatical or spelling errors in a very delimited context). Teachers were firm in their desire not to erode direct writing experiences by using instructional strategies that might enable students to score higher on multiple-choice tests of writing skills but not necessarily make them better writers. They wanted to keep their emphasis on improving content, especially support and elaboration.

Implications for Assessing Elementary School Science

Clearly, ecological and instructional validity issues were paramount in the minds of teachers who supported a direct measure of writing. When applied to science or any other area, this means that tests should be developed to reflect directly that body of knowledge, skills, and understandings that educators feel most strongly about their students' possessing. In science, we must carefully consider the scientific skills and knowledge required for successful performance both in school and out of school in order to develop appropriate assessments. If we want our students to be able to design and carry out experiments, we should be assessing their ability to do so. If we want our students to be able to manipulate scientific apparatus successfully, we should be assessing their ability to do so. If we want our students to understand scientific concepts, we should be assessing their understanding of scientific concepts. If we want to know if our students hold naive conceptions of scientific concepts, we should develop assessments to find out how students conceive of those concepts (see Pechione, et al., 1988). Assessment prototypes currently exist for each of these goals. Compared to multiple-choice tests, many newer forms of assessment require more complex scoring protocols and more intensive time commitments from test administrators.⁴ Many people have asked whether we can afford to build such assessments. Others have responded, "Can we afford not to?"

CURRENT LARGE-SCALE EFFORTS TO USE PERFORMANCE TESTING TO ASSESS SCIENCE

In January 1989, in Tampa, Florida, The Council of Chief State School Officers and the National Science Foundation cosponsored a conference entitled "Alternative Methods of Assessing Science." That meeting demonstrated that performance testing is

taking hold in the United States. Representatives from the International Association for the Evaluation of Educational Achievement (IEA, 1988) and the National Assessment of Educational Progress (NAEP, 1987) described the performance exercises used on the 1986 international study and the Learning by Doing pilot study conducted in the same year. Representatives from the state departments of education of California, Connecticut, Michigan, and New York described their efforts to develop hands-on performance assessments of science. Educators in New York recently completed a monumental effort, administering a hands-on assessment to all 200,000 of the state's fourth-grade students. Connecticut is currently developing a performance assessment based on its Common Core of Learning document, which establishes a standard for an educated high school graduate. Scheduled for implementation in science and mathematics in a sample of high schools during school year 1991-92, these assessments will focus on the integration of knowledge, skills, and attitudes and will employ principles of active and collaborative learning. Students will use higher-order thinking skills to design and carry out field and laboratory investigations; collect, analyze, interpret, and report data; and solve complex problems. This assessment will include exercises of both short and extended duration, as well as development of portfolios, simulations, extended projects, and exhibitions to emphasize students' abilities to engage in meaningful sustained tasks and communicate orally and in writing about their work. One key element will be the assessment of student attitudes, attributes, and interpersonal skills in more authentic, real-world contexts (see Baron, et al., 1989).

At the Tampa meeting, representatives from each of the states and national studies reported on the high level of enthusiasm of the students who participated in the pilot test phases of their projects. Teachers and students not involved felt like they were missing out. Each of the states and the national groups that had used performance assessments of science was confident that performance testing was viable, feasible, appropriate, and valuable given their existing goals for science.

TESTING SCIENCE IN THE 1990s

As we prepare students to assume jobs that don't yet exist and to live in an increasingly more complex and technologically sophisticated world, schools will need to reshape their program goals for science continually. In addition to acquiring a strong base of science knowledge and understandings, students will need a variety of communication, quantification, problem-solving, and

decision-making skills. The rapid rate of change will require students to develop certain dispositions, such as tolerating ambiguity, being persistent, and exploring a variety of options and strategies. Students will need to access relevant information, which is proliferating at a rate that keeps most adults from staying abreast of developments in more than a few selected areas.

What are the implications of this? In the earlier section on writing assessments and validity, I suggested that assessments be in synchrony with the goals of the programs they are assessing. If we assume that our present rate of change will continue to accelerate, we need to ask: What kinds of assessment would be compatible with such a vision? Given that the factual base keeps changing, we might place greater importance on building conceptual understandings of major scientific ideas. One possible guide for an assessment based upon this principle is *Science for All Americans*, produced by the American Association for the Advancement of Science (AAAS, 1989). This Project 2061 report describes the desired residual knowledge of high school students, i.e., "the knowledge, insights and skills that people should possess after the details have faded from memory." Furthermore, if we expect students to be critical thinkers and to evaluate scientific findings in the popular press, then we need to assess students' skills in analyzing scientific arguments and examining generalizations based on data. If we anticipate that students will be working together to solve problems, we should assess their ability to collaborate in problem-solving situations. If we expect students to solve complex multistep problems and demonstrate persistence, then we should include complex problems on assessments. The same holds true for tolerating ambiguity, taking risks, and generating multiple possible solutions to problems and evaluating their merits. If we want students to frame their own questions and attempt to answer them, we will need to find sufficient curricular and assessment opportunities to promote these activities.

Important scientific problems are often ambiguous and poorly structured. They don't present themselves with a set of directions for solving them. They don't announce the appropriate formula or strategy. Rather, they come with a host of irrelevant information and missing data. They require figuring out what information to gather and what procedures to use. They require sustained attention, patience, and flexibility. They require reflection as well as action. They require collaboration. They require self-confidence and a feeling of efficacy on the part of the problem solver. These problem-solving strategies, skills, attitudes, and dispositions in combination with a strong knowledge base are some of the important elements of tomorrow's science curricula and assessments (see Raizen, et al. 1989).

A nation's success in the sciences, technology, and the world economy will rest largely on the ability of its schools and workplaces to be flexible and to adapt their educational programs to the demands of the changing contexts in which tomorrow's scholars and workers will function. Serious efforts to develop multifaceted, performance-based assessments will bring us closer to meeting those requirements. Future efforts to develop appropriate science assessments should be guided by one major criterion: Would we be satisfied to allow our assessments to serve as a manifestation of our science goals? Until the answer is an unqualified yes, we will still have important work to do.

What Has Been Learnt about Assessment from the Work of the APU Science Project?¹

Patricia Murphy

INTRODUCTION

The United Kingdom's Assessment of Performance Unit (APU) is a national monitoring exercise. It arose in response to a general concern about standards in education. Since the APU's inception in 1974, many changes have been instigated in the education process in the UK. The work of the APU influenced these changes, and its unique database both affords insights into the nature of some of the changes and allows critical reflection on them. For example, the work of the APU science project foreshadowed the recent emphasis on incorporating practical problem solving and scientific 'processes' and procedures in school curricula and on assessment initiatives. As part of its assessment framework, the science project staff described aspects of these issues and refined their descriptions through experience and widespread debate. They also developed tests that gave operational and concrete examples of such areas as problem solving. The aim of this paper is to use selected results from the APU science monitoring project to illustrate some of the theoretical and practical implications inherent in recent assessment initiatives in science education.

BACKGROUND

In recent years, here in the U.K., in the U.S.A., and in other countries, there has been widespread concern, particularly at the state level, that science education is failing to meet the challenge of our scientific and technological age. In the U.K., criticism of the science curriculum has led to publication of policy documents on the curriculum (DES, 1985), formation of a curriculum review and development body, production of a plethora of independent curriculum initiatives (which vary in their extent and influence), and radical changes in the assessment methods and objectives advocated for the public examinations given to secondary pupils. These initiatives have been accompanied by an increased onus to assess children's educational achievements in school, which has culminated in the introduction of a national curriculum with national attainment targets (DES, 1987; DES, 1987; National Curriculum

Patricia Murphy is a lecturer in Science and Technology at the School of Education of the Open University. Originally a chemistry teacher, she was a researcher and Deputy Director of the United Kingdom's Assessment of Performance Unit's Science Project at King's College. Publications include research reports, reports for teachers, and articles on assessment and gender issues with particular respect to the science curriculum.

Council, 1989; National Curriculum Council, 1989; National Curriculum Council, 1989).

THE ASSESSMENT OF PERFORMANCE UNIT MONITORING PROGRAM

The APU was set up in the Department of Education and Science in 1974 (DES, 1974). The unit was responsible for conducting a national monitoring exercise, and had as its objectives "to promote the development of methods of assessing and monitoring the achievement of children at school and to seek to identify the incidence of under-achievement." The APU teams of science researchers were established in 1977 at the Centre for Education Studies, King's College, London University, and the Centre for Studies in Science and Mathematics Education, Leeds University. They carried out annual surveys of pupils from 1980-1984. The surveys were conducted at each of three ages—11-, 13-, and 15-years-old—in three countries: England, Wales, and Northern Ireland. A random sample of schools and pupils was drawn and participation was voluntary. Between 12,000 and 16,000 pupils of all abilities and all curriculum backgrounds, from 500-1000 schools, were surveyed at each age, in each year. Since 1985, the project has been conducting in-depth studies of assessment.

In 1988, the United Kingdom passed the Education Reform Act, which established two organizations, the National Curriculum Council and the School Examinations and Assessment Council. Their task is to administer, monitor, and evaluate the school curriculum and its assessment. The work of the APU has become part of the School Examinations and Assessment Council. The APU no longer exists as an independent research body.

THE ASSESSMENT FRAMEWORK FOR SCIENCE

The rationale for the APU science assessment framework reflects the science team's view that science education is an experimental, problem-solving activity involving a complex interaction of demands (cognitive abilities). Accordingly, when pupils engage in such activity, they apply intellectual and practical skills in order to use and develop a body of concepts and knowledge. According to this view, science understanding is a product of scientific conceptual and procedural knowledge. Science education enables children to develop and deploy this understanding appropriately across a range of contents and a variety of tasks. This view of science and of the demands it makes on pupils was the basis for the assessment. The APU produced two publications (Murphy, P and Gott, R., 1984; Harlen, W., et al., 1984) for teachers that describe

the science assessment framework at ages 11, 13 and 15. The intention was to familiarize teachers with the nature of the assessment that schools might be involved in and the type of data available from the surveys.

The first task of the science teams was to review what scientific achievements should and could be measured. The main problem was deciding how to deal with the research team's 'process orientation,' which was a novel feature in science assessment. A secondary problem was to generate assessment tasks that reflected the concern with both process and content and that were appropriate for the full range of pupil abilities within each age group. Finally, the ultimate selection of assessment features had to be educationally valid in the view of teaching professionals, yet still support tests whose results would be reliable and could be reported in a useful form to a wide education audience.

The above description of the project staff's view of science identifies various aspects of children's scientific understanding. The assessment framework is oriented toward one dimension in particular—the scientific processes. This dimension is covered in six broad categories of achievement, the Science Activity categories. The first five of these (Use of Graphical and Symbolic Representations, Use of Apparatus and Measuring Instruments, Observation, Interpretation, and Planning Investigations) represent an analysis of science performance. The sixth category, Performing Investigations, synthesizes the component activities but does not encapsulate all of their characteristics.

The Science Activity categories reflect the focus on the intellectual and practical skills and procedural knowledge within the view of science adopted. The body of science concepts students develop is reflected in a further dimension of the framework, defined in the list of Concepts and Knowledge (DES, 1978). The research team also understood that the content (i.e., the information, object, event, or data in the question) and the context (i.e., the overt question setting—in the laboratory, the home, etc.) would affect the pupils' responses to questions and their perception of the task demands. These two aspects together are captured in the third dimension of the assessment framework.

The framework, translated into questions and tasks, had to integrate the significant dimensions outlined above. The aim of the framework design was to enable question designers to control the degree of interaction between the activity categories and the science concepts. This was to allow some independence in the judgment of performance on these two key facets of science. The independence was not and could not be complete. Essentially, in

the activity categories, there was no requirement to explain specific conceptual understanding recalled in novel situations. However, there was, and always had to be, a need to use conceptual understanding to, for example, observe, interpret, or predict. Similarly, in exercises that required students to apply conceptual understanding to interpret and explain information and data, the process demands were limited to recall, predict, and/or explain.

THE SURVEYS

Two features of the science monitoring program presented particular difficulties from a measurement perspective. These were the need to cover both a broad spectrum of educational objectives in science and a complete range of pupils and their different curriculum experiences. It was essential to select a testing methodology that allowed meaningful comparisons between the performance levels both within different pupil groups (e.g., girls and boys, physics and biology students), and between groups over time. Various attributes of the pupils, the schools, the questions, and their interactions were known to influence science performance. Consequently, the methodology for selecting questions and analyzing performance results needed to take such influences into account. The question selection strategy adopted was domain-sampling. Pupils' performance was analysed within the framework of Generalizability Theory (Johnson and Bell, 1985). The Technical Review of the Project (DES, 1988c) provides details of these and of the test administration.

Three of the categories were tested in the practical mode and three by paper-and-pencil tests (table 1). The written tests included booklets of questions which most 11-year-olds could complete in 45 minutes and most 13- and 15-year-olds in one hour. There were no actual time constraints imposed, however. The tests were administered by teachers who helped pupils with reading and writing.

Two methods of administering practical questions to pupils have been employed (Welford, et al., 1985). In categories 2 and 3, a 'circus' arrangement was used. A typical test session would include 15 questions arranged in nine, eight-minute stations. A pupil would move from one station to another, answering the questions at each. No time constraints were imposed on the primary-age students. The teachers who administered and marked the tests were trained over two days and were assisted by a teacher from the test school. The apparatus was provided in a standard form. In Category 6, Performing Investigations, the administration was one-to-one, and, again, no time constraints were imposed for

TABLE 1. The Science Activity Categories

<u>Category</u>	<u>Sub-Category</u>	<u>Nature of Test</u>
Use of graphical and symbolic representation	Reading information from graphs, tables and charts	Written test
	Representing information as graphs, tables and charts	
Use of apparatus and measuring instruments	Using measuring instruments	Group practical test
	Estimating physical quantities	
	Following instructions for practical work	
Observation	Making and interpreting observations	Group practical test
Interpretation and application	i. Interpreting presented information	Written tests
	ii. Applying:	
	Biology concepts	
	Physics concepts	
Planning of Investigations	Chemistry concepts	Written test
	Planning parts of investigations	
	Planning entire investigations	
Performance of Investigations	Performing entire investigations	Individual practical test

any pupils. The pupils were also given an array of equipment to help them solve their problems.

The data collected ranged from generalized population scores on categories of performance to detailed diagnoses of pupils' errors on individual tasks. In addition, a school questionnaire and a pupil information sheet were developed to collect data on issues thought to influence science performance, e.g., out-of-school hobbies and activities, interests, reading habits. These surveys were developed for each age group. Attitude and interest questionnaires were given to the pupils at the end of a written test package.

REFLECTIONS ON THE FRAMEWORK DEFINITIONS

It is problematic to select the terms by which pupils' attainment in science may be described. No single philosophical or psychological model was found to be an appropriate basis for the assessment exercise and its defined purposes, so none is reflected in the description of the science activity categories. Nor was any hierarchy implied in the list. The activities have often been referred to as synonymous with science processes. At other times, the link is more cautiously stated. Although many generally accepted process terms (e.g., interpreting, observing, and hypothesising) are represented either in the category titles or in specific question descriptors in the categories, the science activities do not define the science processes. Moreover, the operational definitions of the activities include far more components of performance than are normally covered in discussions of the 'processes,' e.g., identifying the status of variables in investigations.

There were various opinions, within the science teams, about the possibility of identifying processes in assessment tasks and the extent of their application across content. This imponderable was resolved by producing tasks that external validators in education agreed fitted a particular subactivity in a category. Each type of task identified was given a description that included what the pupil was given, the expected outcome, and the mode of question response. There was never any intention to define characteristics of the way scientists work. Rather, the concern was to define or represent some of the significant outcomes of science education, either actual or potential.

The identification of process skills and their transferability across content was a matter of exploration and research in the APU program. The key point to consider within the context of the science framework is whether the activities cover the elements needed to develop and use scientific understanding (assuming interaction of process and content in each element).

The APU data are so extensive (see Murphy and Gott, 1984; Harlen, et al., 1984; Welford, et al., 1985; DES, 1981a-1988b; Harlen, 1983; Murphy and Schofield, 1984; Gamble, et al., 1985; Harlen, 1987; and Donnolly, 1988) that only a judicious selection can be offered here. Two areas, one of controversy and one of more general agreement among researchers, have guided this selection. The first focuses on the dichotomy often posed between 'processes' and 'content.' The paper describes results selected to explore the process/content interactions in activities that are here viewed as representing a spectrum of scientific knowledge demands. The second area chosen takes into account the children as agents in learning and assessment situations. Children construct both their knowledge of the world and a sense of themselves in it. These personal constructs determine their subsequent behaviours and responses to tasks. The results included in discussion here refer to the context of tasks, their purposes, and the manner of response open to the pupils. The areas are considered together in three categories of survey results: Observation (Category 3), Planning (Category 5), and Performing Investigations (Category 6).

OBSERVATION

The definition of Observation, within the context of the APU Science assessment framework, has always been problematic. Most of the educators consulted during the early stages of the project regarded observation as a significant scientific activity.

Arguments about the process/content dichotomy often focus on the role and nature of observation in science. Opposing views on observation have been identified in the centuries-old debate about the nature of science and scientific knowledge. The progress and details of the debate are not dealt with here. What is included are some of the main issues that affected the selection of observation tasks in the surveys and the manner in which they were marked.

Since the middle of this century particularly, it has been asserted that observation is theory driven and, therefore, must be a theory-dependent activity. It follows from this that neutral, unprejudiced observations cannot exist. We perceive what we seek, and what we seek is determined by what we know and how we know it.

In contrast, the theory of learning underlying much of present science curricula material often promotes an inductivist view of observation, i.e., that observation is theory neutral. In this view, one observes first and interprets second. Observation is

regarded as a simple, unproblematic process open to all. The theory-driven view challenges this stance on epistemological and psychological grounds. It defines observation as a complex activity, each stage of which is theory dependent and requires children to apply a pre-existing explanatory framework.

The distinction made between observation and inference differs depending on which perspective holds or the extent to which one is orientated toward a particular perspective. The boundary between the two is either clearly demarcated or diffuse and indefinable. For example, if observation is theory driven, what one observes will change as one's knowledge develops. This can lead the observer to attend to more details. When observing a candle burn, for example, increased knowledge might enable one to 'see' the flame in terms of the chemical reaction involved, the degree of carbon formed, the characteristics of the smell, etc. Alternatively, increased knowledge can lead to a reduction in the number of details observed. For example, when classifying objects into metals and nonmetals, the size, shape, and function of the objects may be insignificant observations. The distinction between observation and inference is constantly shifting and has to be judged, and therefore assessed, in the context in which the observations are derived and in light of the purpose for deriving them.

From an inductivist perspective, what is observed is independent of previous experience. This view can have several consequences when translated into practice. For example, the selection and rejection of sensory data would be regarded as a separate process subsequent to observation. When classifying, one would note all features including, for example, the shape and size of the metallic and nonmetallic objects. Rejection of these attributes for classification purposes would occur in the next stage of observation. This secondary process is labeled *inference*. From an inductivist perspective, an inference offered as an observation would be an incorrect response. The theoretical perspective determines the criteria by which observational competence is judged.

Whichever view is held, teaching and assessment that focus on children's observations essentially attempt to understand what they know and how they know it.

Problems Assessing Observation

During the initial review of existing test instruments, the science teams found very few suitable questions, whether in use in schools or in assessment literature. The first problem for the team prior to generating a question bank was to define scientific observation. Table 2 describes the question types finally agreed on

(DES, 1989a). A range of content, contexts, and resources was used within any one type. Various modes of presentation, operation, and response were also employed.

TABLE 2. A Description of Types of Questions Surveyed

<u>Given</u>	<u>Task</u>
Objects or photographs	Group objects into self-defined classes or identify the rules used to classify the objects and add further objects to classes.
Objects, photographs, or events	Describe similarities and differences.
An event	Make a record of change.
An object and a range of drawings	Select the matching drawing.
Events	Make a record of observations and either give, or select, an appropriate explanation.
Objects, photographs or events	Make a note of differences and make, or select, a prediction consistent with observed data.
Events	Make a record of changes, and either make a prediction or identify a pattern in the observed changes.

The science teams had difficulty translating the activity of 'observation' into questions and mark schemes that did not conflict with the overarching view that the category was essentially about 'process' and not dependent on the recall and application of science concepts. This was a problem at each age level. We had to assess pupils of all abilities from the age of 11 through 15. Since we viewed observation as a theory-driven activity, we recognized that the same task could vary in the demands made of pupils from one age to another and one pupil to another. Yet we had to assign scores unambiguously. Some of the difficulties we experienced in assessing pupils' observational competence are illustrated next, through examples of tasks and our problems developing them and interpreting pupils' performance on them.

Observation Tasks

The purpose given in an assessment task is fundamental since it determines the appropriate knowledge to draw on. The activities in the assessment framework had to be structured and fragmented for analysis and reporting. This feature of assessment tasks makes it very difficult to incorporate a purpose other than the task itself. But if a clear scientific purpose cannot be established, how can pupils judge what is relevant and how can their responses be scored? This was particularly difficult in observation questions that asked pupils to compare resources. One such task asked children to describe a number of differences and similarities between two pieces of bark, one from a silver birch and the other from an elm tree.

We scored any correct observation and made no distinction, in terms of scores achieved, for the different theories applied. The number of observations recorded was credited in the scoring system. Furthermore, statements judged to be purely inferential were not scored. Thus, no score was given if pupils describing similarities and differences between crab shells noted that they "live on the beach" or "they come from the sea." The pupils had to understand and operate within the intended assessment purpose—observing shells, not whole crabs (living or dead). Inferences about crabs' way of life or even about structures not visible from the shells were all judged to be invalid. But is this realistic? We want pupils to use their knowledge and to make deductions in answering certain questions and to know that this is not appropriate in others. We rarely supply guidelines to this effect, in either assessment or teaching situations. The decisions we made for the APU assessments were the result of trying to produce reliable mark schemes, and the proportion of responses considered to be purely inferential was low (2 percent). Thus, for us, the dilemma was not in the end a validity problem.

The biggest threat to the validity of these questions lay in the lack of cues given to pupils about how to observe the resources. Without purpose, the pupils have to define and impose their own 'model' of the presented situation. Depending on the pupils' background experience and knowledge, different resources will cue different 'models.' One important consequence will be the resulting perception of what is relevant and noteworthy. Certain overtly scientific resources may well focus the pupils into specific perceptions of relevance. For example, when observing seeds dispersed by the wind, some pupils may focus only on such related features as shape and mass, rather than describe the colors, surface patterns, and texture of the seeds. These pupils could actually

achieve lower scores on mark schemes based on an atheoretical view of observation, which give credit for the number of observations made rather than the type. This ambiguity arises because of the need to distinguish between process-led mark schemes and concept-led mark schemes in attempting to allocate scores to separate dimensions of scientific achievement.

Establishing a purpose and using appropriate cues is particularly fraught with difficulty for primary-school children. If science has not been identified as a discipline, which is generally the case, then the domain does not exist to be cued. However, open tasks do serve a useful assessment purpose if the interest is to establish the characteristics of novice observers and to diagnose the difficulties they may experience in scientific observation. The problem with the tasks arises in marking children's responses and defining criteria for observational competence. Ideally, the mark schemes should take account of valid alternative and qualitatively different perceptions of resources that do not exclude either the process-led or the concept-led response but differentiate between them. The criteria for scoring need to relate to: the type of attributes identified as relevant, e.g., structural or nonstructural variables; science concept dependent or otherwise; the theoretical model of the task developed, e.g., within science or outside of it; and the precision of expression employed. Mark schemes that reflected these suggestions would define part of the continuum of scientific knowledge demands called on in the domain of observation.

Some characteristics of this continuum are demonstrated in 11-year-old pupils' responses (figure 1) to one observation task, 'Sound Box.' The 'Sound Box' task required pupils to identify relevant variables and to relate sound frequency with string diameter. Students were given a rectangular block of wood with wires of the same length but different diameters stretched between pegs. They were told that this musical instrument was like a guitar, then asked to answer the following questions: a) What do you notice that is different about each of the strings? b) Now pluck each string one at a time. Describe how the differences in each string affect the sound it makes.

The first group of pupils' responses gives some indication of the alternative questions perceived by pupils, i.e., alternative to that intended by the assessor. This is the result of the pupils' engaging with the resource first and foremost and not with the assessor's cues to the task.

The second set of responses (i) - (iv) illustrates some of the difficulties pupils experience in describing a relationship they have

FIGURE 1. Examples of Pupils' Responses to 'Sound Box'

1. Alternative Question Answered

it is harder to pull)

It sounds like a door bell.

Number three is the loudest of them all.

it echoes when you play it.

2. Both Variables Related

(i) Direction of Relationship Given

The thinner the string the higher the note.

The thinner the string the higher the sound and the thicker the string the lower the sound.

The thicker the string gets the deeper the sound.

(ii) Not all Observations Related

The thickest string makes the lowest noise
The thinnest string makes the highest noise

(iii) No Direction Given

The different thicknesses make different sounds

(iv) Observations Specified and Not Generalised

The first one makes a high tone the second one makes a slightly lower noise, the third makes a deep sound and the fourth makes a slightly low sound.

3. One Variable Only

They are all different sounds

It gets higher and lower and deeper sound

4. Discrimination And Language Difficulties

The top one sounds weakish and
the 3rd one sounds higher

The first one make dull more second a bit
lighter third an Indian sound fourth a King)
chilling sound.

3 of them are here and 1 is low

The thinner the string the sharper the
sound.

5. Additional Variables Considered

The big thick silver one make a lower
beat than the others. The thin silver one
is the highest note. The thick copper one
is the next lowest sound. The thin copper
one is the next highest so there are two
high ones and two low ones

the thick copper wire makes a dull sound...
and the thinner one makes a sharp sound.
the thick metal wire makes a duller sound than the
copper. The thin metal makes a sharper sound than
the copper one.

the thin metal one is higher than
the copper one but the thick copper
one is higher than the thick metal

6. Alternative Variables Considered

The very thick string vibrates.

the tighter the strings are
the lower the sound gets.

The strings are different types
of thing so they vibrate differently
the thickness affects this as well

The tighter the string the higher the note

observed. The third set of responses refers to the number of variables identified as relevant. This type of partial response is often related to a view pupils hold of the dominant variable in a presented situation, which leads them to reformulate the task they have been given.

The fourth group is concerned with pupils' ability to discriminate between sensory data and to express these discriminations. The issue of alternative views of relevance is highlighted in the fifth and sixth groups. If, as in the fifth group, pupils do not discard those variables considered irrelevant in the assessor's model of the situation, the difficulty of the task is often increased and simple relationships remain 'obscure.' Pupils' responses in the sixth set suggest that they imposed an alternative theoretical model of the task. In some instances, this can lead to absent data being generated; in others, to the perception of a theoretically more complex relationship.

This range of responses can be found at each of the ages surveyed. On scientific grounds, it can be said that the quality of observation improves as pupils get older but this is by no means true for all pupils.

CLASSIFICATION

Tasks of this type (figure 2) have been used with both everyday and science content. Pupils' performance when identifying the rationale upon which a classification is based (figure 2, part a) is generally high for an everyday content. The classifications to be identified were related to the everyday functions or known characteristics of the resource. Hence, the purpose and the relevance of the groupings could be assumed. There was also no potential for ambiguity arising from competing theories about the resource.

In questions that were identical in presentation, but where the resource was judged to have scientifically relevant attributes, pupils' performance was much lower. We found, for example, that when different colored liquids were grouped by viscosity, more than 40 percent of 13-year-olds did not observe this attribute, or, if they did, they discarded it as irrelevant. Many of these pupils imposed their own 'everyday' model of the resource and observed accordingly. Thus, the unnamed, firmly sealed liquids were classed as "antiseptics," "disinfectants," "cleaning things," or "strongly smelling." This is a clear example of theory-driven observation, matching what is seen to what is known. The proportion of 11-year-olds using everyday classifications instead of scientific ones was even higher. In all cases, the everyday classifications required

FIGURE 2. Classification Problems

Example a: Buttons

You have been given a collection of buttons.

The buttons have been divided into three groups, P, Q, and R.

Look carefully at the buttons. You may pick them up and handle them.

- a) Decide how the buttons have been grouped and write this down in the spaces below:

Group P are all

Group Q are all

Group R are all

- b) Now think of another way of sorting the buttons into three new groups, X, Y, and Z, of equal size. You may pick them up and move them about.

Write down the three groups you choose in the spaces below:

Group X are all

Group Y are all

Group Z are all

Put the buttons back into groups P, Q, and R when you have finished.

Example b: Bones

You have been given a collection of bones all labelled with letters.

These bones are from the back bones of different animals.

Put the bones into three groups making sure that there is something the same about all the ones in each group.

Write the letters of the bones in your groups in the columns below:

Group 1	Group 2	Group 3

the children to make large inferential leaps unsupported by the observations available to them.

When asked to generate exclusive groups from a collection of presented objects, 11-year-olds commonly produced classification systems in which the groups overlapped. This response is also apparent in classroom situations. We cannot assume, therefore, that pupils understand the nature and purpose of classification systems. This lack of understanding of what the procedure of classifying is about, not the associated science concepts, is problematic for many.

Questions that required pupils both to identify classification criteria and reclassify (figure 2, part a) used a very particular type of preselected resource—for example, a collection of objects that could be classified into three equal groups by two separate attributes, such as man-made/natural; metallic/nonmetallic. In addition, redundant distractor variables such as color and shape were deliberately included. The pupils, therefore, had to operate within a closed, premodeled system in which they had to identify both the model and the variables. Many of the pupils' responses indicated a failure to operate within the presented model. Other questions used in the surveys had the same demand, to classify, but the presentation and response format was open (figure 2, part b). Most pupils at 13 and 15 and many at 11 answered these questions well and demonstrated that they could classify.

What have we learned about this type of task? When pupils are classifying content they select relevant features of it and discard others. Moreover, to use their science concepts appropriately in order to make observations, they must understand the task itself and know something of the resource to be classified. They need to understand the procedure and its purpose. These are all prerequisites of successful scientific classification. We cannot assume that the general activity of classification is unproblematic for pupils. Many older pupils do not incorporate developing knowledge into their existing classification systems to establish more sophisticated ones. For example, most 15-year-olds continue to regard metals in the light of their everyday observable properties, e.g., hardness and strength (Donnolly, 1988). Consequently, their classifications remain very similar to those of 11-year-olds despite their additional experiences. This suggests that the learning situations they are involved in do not enable them to link their knowledge of metals (e.g., car bodies, cutlery, jewelry) to the strips and blocks they encounter in the classroom and school laboratory.

DISCUSSION

In our exploration of the observation domain, we have found immense complexity. This complexity has two clear sources: the multiple nature of the task demands and the wide range of purposes for which observation is deployed.

Pupils have conceptual difficulties with questions that require them to both use knowledge to select relevant observations and explain the use. Pupils responding to tasks that require descriptions or comparisons of objects and events do not often offer inferences. Pupils more often note descriptive rather than abstract observations (e.g., color rather than structure; shape rather than composition). Also, pupils' different curriculum backgrounds promote qualitatively different responses. Pupils studying biology achieved higher scores on this test than those studying physics or chemistry.

Test questions that require students to apply conceptual understanding to make appropriate observations in tasks can be manipulated to emphasize use of knowledge rather than recall of explanations. However, in many tasks the knowledge applied cannot be predetermined. The pupil judges what the task means and what an appropriate response is. Very often, pupils will apply everyday theories in preference to scientific knowledge.

The overwhelming majority of pupils make correct observations. Yet the nature of the observations made and the range of senses used to characterize objects and events is limited from a scientific perspective. Pupils need experience with a wide range of content, and purposes for considering it, if they are to begin to explore and develop scientific concepts.

Pupils' performance on, and engagement with, a task is affected by the complexity of the resource or event. The number and nature of the variables to be encoded are not trivial features of science observation tasks. The more variables and the more abstract their nature, the lower the performance. Significantly, children may 'see' more in events than the adult, theory-driven observer who focuses, often unconsciously, on just those variables to be related.

Finally, we cannot assume that certain activities are unproblematic for all ages or for pupils of all abilities. If the concept of a 'group' and its function is not understood, then classification as a procedure cannot be carried out successfully.

INVESTIGATIONS--THE APU PERSPECTIVE

A central part of the assessment of scientific performance is based on pupils' responses to practical science investigations

(Murphy and Schofield, 1984) (table 1, category 6). The APU science assessment framework (Murphy and Gott, 1984) reflects the view that the use and further development of scientific understanding is achieved through theoretical and experimental problem-solving activities. The APU research in this test category highlights the relationship between the procedures and content of science within activities, pupils' understanding of it, and some of the factors that mediate their understanding and, hence, their ability to transfer this understanding to new situations. The assessment values the process of investigating, not just the outcome. It also details pupils' common strategies and errors. A summary of the rationale for the categories and the main survey, and associated research findings, follows. For ease of communication, the description provided is largely qualitative but refers only to significant and characteristic responses of pupils in all age groups.

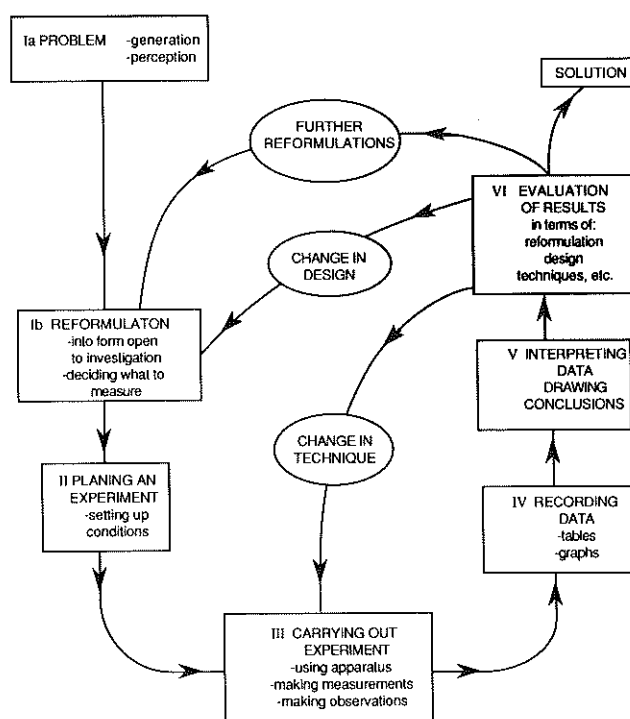
For the assessment, a problem was defined (Gott and Murphy, 1987) as a task for which the pupil's success depended on more than recall of a well-learned solution. Practical investigations were viewed as one type of problem. The development of the assessment tasks and analysis methods was based on a particular model of how pupils respond. The evolution of this model is detailed elsewhere (Murphy, 1987). According to this model, pupils must access the pool of knowledge available to them before they can perceive that there is indeed a problem to solve. The student must then generate a procedural strategy; the pupil's particular conceptualization of the problem determines the procedures he or she perceives as appropriate. Pupils' interactions with features of the tasks mediate both the problem perceived and the strategy developed.

To aid the reporting and analysis of the results, we made the following distinction between the types of scientific demands that investigations place on pupils: Pupils need both generalized conceptual knowledge (i.e., the theoretical constructs that relate and give meaning to the observed properties of objects and phenomena) and generalized procedural knowledge of the means by which the conceptual knowledge is deployed and developed. Little was known about pupils' procedural understanding in science, so the assessment initially focused on this aspect of pupils' performance. Students needed only a tacit understanding of such concepts as mass, volume, and thermal conduction to understand the variables in the investigations used.

The investigations, 36 in all, were generally concerned with the relationships between variables and their effects. The investigations covered a range of purposes, contents, and settings, and the

complexity and type of procedures needed to solve the investigations varied (Murphy, 1987). A general mark scheme of procedural activities was developed (figure 3). The figure depicts part of a spiral of activities. In each loop of the spiral, as defined by figure 3, all of the stages, or a selection of them, are involved in a variety of ways as pupils make decisions, refine their perception of the problem, and collect and evaluate sensory input. Initially, pupils will engage at a theoretical level as they consider the problem, the type of solution needed, and the possible nature and scale of the data collection.

FIGURE 3. A General Mark Scheme



The scheme functioned as a general heuristic in that it represented a strategy, independent of question content, that helps scientific problem solvers approach and organize their resources—e.g., knowledge, experience, equipment, information—when solving problems. We used checklists to collect the data about pupils' actions. The checklists took account of the different pathways pupils take to solve problems. The checklists allowed assessors to grade adequacy of performance for the main elements in the cyclical model. We used a 'strategy sieve' to analyze the results.

fine mesh of stringent requirements used for the first sifting allowed only those pupils who used the ideal strategy to fall through. This was followed by a systematic relaxation of the specifications. The application of each gauge of mesh allowed more and more pupils to fall through until all pupils were taken into account. In this way, a full description of pupils' performance was achieved. The general analysis features included in the check-lists are shown in table 3.

TABLE 3. Analysis Features

Problem perception	<p>Defining the status of:</p> <p>the independent variable, i.e. that to be varied systematically in the investigation</p> <p>the dependent variable, i.e. the variable affected</p>
Problem reformulation	<p>how to vary the independent variable</p> <p>how to judge the effect on the dependent variable</p> <p>identification of variables whose effect must be kept constant (control variables)</p>
Planning and carrying out	<p>Setting up:</p> <p>the test of the independent variable</p> <p>the measurement of the dependent variable</p> <p>the control of other variables.</p> <p>Taking account of:</p> <p>scale for the quantities of variables</p> <p>range of readings</p> <p>(both of which depend on the nature of the variables, the type of answer necessary, and the measuring instruments available)</p>
Recording and interpreting	<p>developing a strategy for sifting complex multivariate data</p> <p>transforming data from one form to another (here account has to be taken of the whole task and the type of solution appropriate to it)</p>
Evaluation by the pupil	<p>assessing the data against the demands of the question and the answer to be derived</p>

Interviews were conducted to find out more about pupils' decision making and views of the problems. The typical sample size was 500, although this was reduced in 1984. The test methodology and investigations were extensively trialed before the surveys (DES, 1988a; DES, 1989; DES, 1988b).

SOME FINDINGS

Most pupils were able to apply scientific understanding successfully when carrying out investigations of phenomena they were unable to explain adequately in the written tests. The potential to define their own model of the problem seemed to contribute to the pupils' apparent confidence and initial success in this assessment situation. A sense of control over their own ideas and subsequent decisions pervaded the comments pupils made about their liking for the investigations. This was true of pupils of all ages and abilities and despite the requirement for them to work on their own under scrutiny. The vast majority of pupils also showed considerable commitment to their investigations and worked for extended periods of time, generally longer than 20 minutes. The assessors who administered the assessment stressed that they were interested in what pupils chose to do, thus there was no 'challenge' to students' models.

Pupils in all age groups successfully set up *minimum designs* for a range of investigations. However, many factors affected pupils' ability to reformulate and pursue problems.

Pupils' Perceptions of Problems--The Effect of Cues

The content and the setting within which the investigation is presented are significant cues for the pupil. For example, more girls than boys in the survey sample were anxious about their ability to work with real equipment. Many of these pupils will reject investigations involving a range of apparatus because they believe that they do not have the necessary skills to tackle them. The practical demands are perceived to be the 'problem.'

Girls at each age tended to reject investigations that included content related to electricity. They expressed the belief that they did not have the necessary knowledge, regardless of whether the investigation required them to apply any specific understanding of electricity. The girls reformulated the problem as outside their domain of competence.

Some boys rejected investigations with an obvious domestic content, such as choosing the most suitable floor surface for a kitchen. Here, the problem was perceived to be 'outside' of science and therefore not appropriately addressed by their knowledge.

When the content of an investigation is familiar to the pupils, they often fail to 'see' that there is a problem for them to solve. For example, some pupils refused to investigate the environmental conditions preferred by woodlice. They 'knew' that woodlice like wood, or that they live in the garden, because "they've seen them." In this situation, the pupil does not perceive that a problem exists.

Another effect of familiarity with the content is that an alternative problem can be cued. For example, 13- and 15-year-old pupils were asked to investigate the effect of the length of a spring and the diameter of the coil on the rate at which the spring oscillates. The springs cued the Hooke's law experiment for some pupils who proceeded to measure the extension and not the rate of oscillation. As science becomes increasingly a part of the primary school curriculum in England and Wales, similar effects will be noticed in assessments of 11-year-olds.

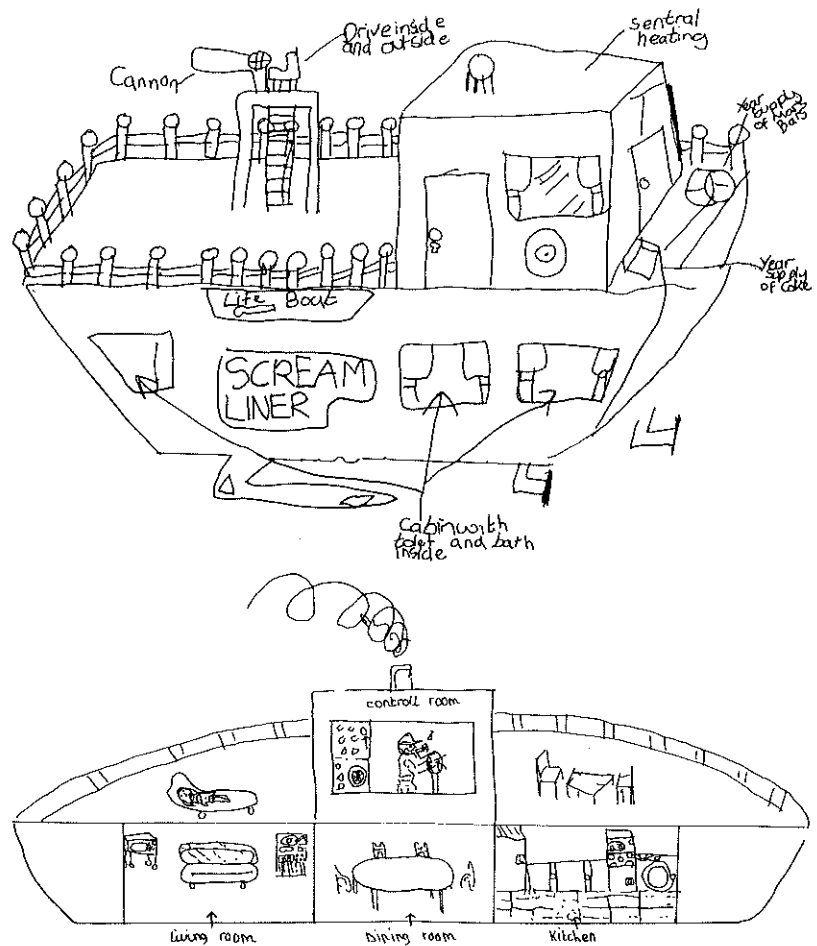
Pupils' ability to construct meaning in a task depends on their experience, scientific understanding, and views of science and of themselves. This is true even for investigations where the demands for scientific knowledge are reduced. For example, 13-year-olds were given the same problem in two contexts, everyday and scientific (DES, 1989). In the first, they had to find the effect of temperature and size on the rate at which a chemical dissolved. The equipment provided included beakers, measuring cylinders, spatulas, etc. In the second, they had to find the effect of temperature and size on the rate at which a brand of candy dissolved. The equipment provided included plastic cups, measuring jugs, teaspoons, etc. About 50 percent of the pupils carried out an effective quantitative approach in the science setting compared with 25 percent for the everyday investigation. Pupils explained that an everyday problem does not require a 'scientific' solution, and they tended not to control variables or take measurements. The context dominated the actual problem perceived and dictated the solution. This occurs to an even greater extent at the primary ages.

Another way in which the context can alter the problem set occurs more for girls than boys. Several of the investigations were set in the context of a human dilemma. One problem concerned survival of a person stranded on a mountainside. Pupils were asked to "find out which material would keep a person warmer." They were even advised to wrap cans of hot water with the materials provided. During the investigation, some girls were observed dipping the materials in water, blowing air through them, and even making prototype coats. For these girls, the context overrode the

cued investigation. Thus, they investigated how porous or water-proof the material was and whether it was indeed suitable for making a coat. The girls' problems were frequently more complex and required sophisticated procedural strategies to solve. However, all too often such behavior is labelled as "off-task" and "incorrect." Examples of this kind were found in all the age groups surveyed.

This result was assessed further through open-ended tasks, such as designing a new vehicle or a boat to go around the world, that were given to pupils aged 8 to 15. Striking and consistent differences in the designs of girls and boys reflected the concerns that they are encouraged to pursue (Murphy and Moon, 1990). Most of the girls' designs, unlike the boys', dealt in detail with the daily needs of people (figure 4).

FIGURE 4. Examples of a Boy's and a Girl's Drawings of Ships



When asked to investigate an aspect of their initial design, the majority of the boys did so with little trouble. Most girls remained with the complex human problem they had perceived. It is not a simple matter to reject a perception of human need in order to focus on a more artificial concern, such as the learning of a specific subject outcome or the assessment of the same. Furthermore, an ability to do so may reflect only a limited and uncritical view of problems and problem-solving strategies.

Pupils' Procedural Understanding

The procedural strategies pupils develop are also dependent on their conceptual understanding. Metaphors and analogies play an important role in forging links between pupils' existing and developing knowledge. They also enable pupils to understand the variables within investigations. When asked to investigate the effect that the height of water in a container had on the rate of flow from it, few 13-year-olds were able to grasp the problem. Yet when given an 'image' of the variables in the context of a tea urn gradually emptying and a queue of people simultaneously growing, the majority of pupils 'saw' the problem. Many pupils have difficulty understanding various 'rate' measures, yet most 11-year-olds in primary schools can handle rates of cooling or dissolving because they can draw on common, everyday experiences of these effects.

In order to identify the variable to be changed in an investigation, pupils must first understand that a variable can take any value and that the value can be changed and measured. A significant proportion of 11- and 13-year-old pupils have difficulty conceptualizing a range of variables in this way. They either fail to understand how the variable can be manipulated or they operationalize it in terms of their everyday understanding. For example, temperature is translated into hot and cold, i.e., two extreme values only. If, however, the independent variable is operationalized when presented (e.g., 'type of material' is presented to the pupils as pieces of blanket and plastic), the requirement to understand the variable (e.g., 'type of material') is removed. In these circumstances, more than 90 percent of pupils in all age groups successfully test the variable. For the reasons described, an investigation involving a continuous, independent variable is too difficult for most 11-year-olds, many 13-year-olds, and some 15-year-olds.

Most 13-year-olds can deal with two independent variables if the variables can be treated as discontinuous, i.e., as specific quantities, such as hot/cold or whole/halves, rather than as values on a scale that the pupils must select. More 11-year-olds suffer

from information overload in these situations. If pupils must manipulate the variables to establish conditions in which both are varied systematically, then about 20-30 percent of pupils at age 13 have conceptual difficulties and an even larger proportion at age 11.

Most pupils in all age groups surveyed appear to control variables in investigations successfully. However, this measured success can be misleading. Some pupils control variables unconsciously while carrying out a 'neat' investigation. It is not a tactical decision. This was found more often at the younger ages tested. However, the majority of pupils' actions are thoughtful, and their ability to identify control variables depends on their understanding the variables' effects. When setting up two heat sources to compare the thermal conduction of materials, 70 percent of 15-year-olds controlled the volume of hot water, but only 40 percent controlled the temperature. The dependence of procedural success on conceptual understanding means that pupils' actions are easy to misinterpret.

When asked to compare thermal conduction, some pupils placed the test material over the top of cans of hot water, like jampot covers. This left the sides of the cans unevenly exposed, and could be coded as a failure to control the amount of material and the cooling conditions. However, the pupils explained that 'heat' escapes from water only as steam, so only the top of the can needed to be covered. The pupils were competently controlling the variables they understood to have an effect.

Measurement Strategies

The purpose of the science problems students are given varies. Often, the purpose is to collect variable-based data from which patterns and predictions can be generated and concepts developed. At other times, the focus is on pupils using their knowledge to solve problems. During this process, pupils develop understanding of the nature of scientific procedures and evidence and of the various problem solutions that can be derived. A failure to address this latter aspect of pupils' learning will limit the success of the former. In problem-solving activities, pupils have to decide what procedures are necessary to reach the required solution. This includes deciding when to measure, what to measure, and how.

Across investigations, about 20 percent of 13- and 15-year-old pupils did not see the need for, or advantage of, measurements. This figure is about 50 percent at age 11. To decide what to measure, pupils have to apply knowledge to interpret the meaning of the terms used in investigations. To find out "which of three kitchen

towels holds the most water,” pupils have to translate ‘hold’ into a variable that can be measured. Some 13-year-olds and over 30 percent of 11-year-olds had difficulty with this. A common translation was the everyday one: “How much will it carry?” This led pupils to make bags of the towels and to attempt to measure ‘carrying potential.’ If the word *absorb* was used instead of *hold*, many more pupils failed to understand the term. Replacing *hold* with *soak* ensured that most pupils understood the meaning. However, *soak* triggered an everyday image that cued an everyday response. Hence, in this situation many pupils failed to ‘see’ the need for quantified data.

Pupils commonly fail to identify the dependent variable correctly in investigations because their understanding conflicts with the scientifically acceptable one. In the investigation in which students were asked to judge the thermal conduction of materials, 15 percent of 13-year-olds thought the “heat was held by the material.” Their procedural strategies reflected this understanding. They wrapped the material around thermometers or cans of cold water and waited for the temperature to rise. Alternatively, they judged the ‘warmth’ for themselves. More than twice as many 11-year-olds hold this type of understanding. In animal behavior investigations, most pupils in all age groups attributed human patterns of behavior to the animals. This perception lead them to judge animal preference, the dependent variable, by the ‘moods’ manifested by the animals. Thus, they looked for signs of happiness or discontent. This interpretation affects their whole design, and, consequently, the performance of 13- and 15-year-old pupils is very similar to that of 11-year-olds: Only 10-15 percent achieved success.

A common response of 11-year-olds when dealing with ‘rate’ measures is to hold one variable fixed rather than attempt to judge the relationship between the two variables. For example, when judging the rate of swing of boards or pendulums, more than a quarter of these students will measure the time to the end of swing. This is a common translation of ‘rate’ for primary school pupils.

The pupils who decide that a quantified approach is appropriate have then to understand how to affect the measurements. This understanding is far from simple. Pupils have to predict the range of the potential data. Their ability to do so depends on their estimating skills and theoretical understanding of the solution. The surveys show that all pupils have difficulty estimating quantities with any degree of accuracy. For example, an 11-year-old predicted in a planning question that a rabbit would gain about 2 stone

(28 lbs) every few weeks. This degree of inaccuracy is not unusual. Yet a growing rabbit is much more familiar than examples included in typical science investigations. Even older pupils can estimate only a few simple variables with any confidence. This lack of knowledge affects pupils' decisions about the frequency with which to take readings and with what instruments.

We teach pupils graphing skills and expect them to know when a graph is the best form of data representation. This knowledge alters pupils' procedural strategies as they decide how many readings to take, at what intervals. The survey results indicate that the majority of 15-year-olds have mastered graphing skills, yet few (20 percent) choose to plot a graph when necessary. Similarly, most 11-year-olds can construct and use tables and bar charts yet less than 6 percent use them to organize the data they collect during investigations. Most pupils successfully read scales of instruments, though their success depends on the type of scale and the mathematical operations involved. However, choosing an appropriate instrument requires an understanding of scale and associated errors. Few 13- or 15-year-olds demonstrate this understanding. The majority do not use the best instruments available or take account of scale in their designs—these are decisions they rarely face in class.

'Knowing' certain procedures can actually inhibit pupils' approaches to investigations. This is a necessary phase in moving from an everyday world view to a scientific view. In the investigation of which kitchen towel held the most water, 8-year-olds automatically tore off whole pieces of towel at the perforations. This unconscious choice of scale allowed them to obtain readings on the insensitive balance provided. Eleven-year-olds, in developing a thoughtful, scientific strategy, noticed that the pieces were different sizes and carefully measured equal-sized pieces. These did not register on the balance. The older pupils modified their strategy, which got progressively worse, trying to deal with the difficult ideas of scale and control.

Planning Investigations

The planning category (table 1, category 5) represented an assessment of pupils' recall and application of known procedures. The category was assessed at each age level, in four-to-five hours of testing covering 60-80 different planning questions. Pupils were required to generate complete plans or specified parts of plans for a very wide range of practical investigations. A brief review of pupils' performance on planning investigations is included here to further illuminate aspects of pupils' practical problem-solving

performance and the factors that affect their success. In particular, the results for the same investigations assessed in both categories 5 and 6 (table 1) are described.

In their written plan of whole investigations, pupils were generally more divergent in their overall approach. The lack of practical cues resulted in students generating many more alternative investigations. In devising these alternative investigations, students drew on external information that related to the specific content and context of the problem situation presented but ignored the actual investigation posed. The pupils' alternatives were generally very creative but usually impractical and impossible to carry out.

Pupils rarely mentioned any control variables in their plans. Yet their ability to control variables in the practical tests was very high. This low performance on the written plans occurred in all three age groups tested. Similarly, pupils' plans included very little detail concerning the manner of carrying out measurements. At most, pupils mentioned a final measurement irrespective of the need for a base-line measure or whether a relationship was being investigated.

When a picture of the range of apparatus used in the practical test was included in the written plan question, performance was generally elevated. Pupils were cued more into adopting a scientific approach to the investigation. Pupils' written accounts after carrying out the practical investigations were reasonable representations of what they had done. Thus, the process of writing itself does not seem to be an obstacle. Rather, students were stymied by the need to carry out 'thought' experiments in a nonpractical context without any stimulus for further thought.

In the practical investigations, many pupils were unable to develop a full strategy and/or retain it. We thought that the pupils were possibly being given too much to handle in the course of an investigation, and that questions that asked them to plan just specified parts of an investigation would reduce this burden. However, performance for every aspect identified in the analysis was lower on questions that asked students to plan only part of an investigation. Evidently, the ability to 'see' the whole task and to obtain concrete feedback reduces the demand on the pupils to retain a strategic view in their heads. Thus, paper-and-pencil questions on aspects of investigations increase rather than decrease the demand to have an abstract, strategic view for tackling variable-based problems. This is surprising because a structured question form provides many clues and cues absent in an open form. The

fact that these do not appear to help confirms again that the overall strategy, not detailed tactics, is critical.

Although the pattern of performance across the various procedural demands specified in the planning parts of questions was lower, it matched that found in the practical context. Performance was lowest on questions that asked pupils to identify what to measure in a problem and to describe a strategy for the measurement. Less than half the pupils obtained any marks on these questions. At age 11, the majority of pupils were unable to specify even the quantities to be measured, let alone any strategy for their measurement. The results of the practical assessment suggested that a pupil has to have the concept of a procedure in order to deploy it appropriately. Many 11- and 13-year-old pupils responded to questions about what and how to take measurements in an investigation by quoting either a measurement instrument or the units. For example, children were asked what measurements they would take to find out which cup is the best for keeping soup warm. Common responses included "temperature," "°C," or "thermometer."

DISCUSSION

The surveys indicate that most pupils possess fragmentary scientific knowledge that is usable only in narrow, context and content specific, instances. Similarly, pupils' data representation and measurement skills are divorced from an understanding of the nature of scientific relationships or evidence and, thus, from their applications. Pupils apply the simplest strategies across investigations, regardless of the solutions required. We clearly need to reconsider the nature of the learning opportunities afforded to pupils and the demands in activities that we may consider trivial and unproblematic. At the very least, pupils who are expected to apply procedural knowledge must have concepts of number, scale, and variables. These concepts have to be:

- integrated with an understanding of scientific relationships and evidence;
- applied in the light of specific science concepts;
- used with techniques of instrumentation that assume an understanding of measurement error.

Pupils need experience with a range of problems that highlight different conceptual demands and provide opportunities to generate a repertoire of strategies. The investigations described in this paper are useful because they allow pupils to engage at a

level appropriate to, and indicative of, their understanding. Thus, they are motivating and afford insights into the nature of pupils' scientific understanding. They also provide pupils with feedback about their decision making and its effects.

Students had a higher success rate for the Performing Investigation category than for most other forms of assessment used in the surveys. Pupils' performance in planning and carrying out investigations has been compared with and without apparatus for the same problem. The features that seemed to characterize pupils' success at investigations include the ability to interact in a practical context, respond in action, interact with a 'whole' task, and develop and work within a 'personal model' of a problem.

SUMMARY AND IMPLICATIONS

The survey results confirm that science performance is always an outcome of a process-content interaction. Moreover, there appears as yet to be no identifiable group of pupils who can operate on a process at the same level across a range of content. Emphasizing the processes and procedures in science and reducing the requirement to recall specific concepts has allowed us to understand a wide range of cognitive demands that are significant in scientific activity. We also are increasingly aware of how these demands affect pupils in all age groups. The model of pupil-task interactions (Murphy, 1987) allows a great deal to be said about the appropriateness of tasks, and of task presentation, for assessment purposes. The APU research findings enable us to set the level of demand within an investigation quite accurately. If one feature of an investigation is intentionally burdensome, we can reduce the demands of others while still maintaining the validity of a real task. The model can therefore be used as a powerful generator of diagnostic tasks. It can help teachers both to guide pupils in the difficulties they experience learning science and to assess their progress.

A review of the APU results on the category tests reveals some characteristic features of student performance that have direct implications for assessment. For example, pupils can read and construct coordinate data forms appropriate to their age. However, their success is limited by their understanding of the specific variables in the data, the type of scale to be manipulated, and the mathematical operations involved. Pupils can also successfully read values from a range of instruments as long as the scale involved is not complex. If these factors are not separated out, performance levels on graphing skills, etc., can be raised or lowered by means unrelated to the criterion being assessed.

On the other hand, few pupils label graph axes, name graphs, or indicate the units of quantities measured or represented. This raises the problem of the arbitrary divides (e.g., the distinction between using graphs [table 1, category 1] and interpreting graphs [table 1, category 4 [i]], etc.) constructed between assessment criteria in category- or profile-based assessment. The naming of graphs and graph axes (assessed in category 1) assumes a complete understanding of both the algorithm and the presented data. Such an understanding comes close to what we assess in other categories of science performance, e.g., data interpretation (category 4i) and planning (category 5). It is often necessary, therefore, to use questions that span several categories of activities. By contrasting pupils' performance on these questions with their performance on questions covering individual categories, assessors can obtain a better understanding of pupils' achievements and difficulties. It is a rare assessment model that allows this approach. However, without flexibility of this kind, we risk providing misinformation about pupils' achievements or information that has limited interpretability.

Although pupils have mastered the techniques of graphing and measuring, few knew when or why to use them when carrying out investigations. Half of the 11-year-olds assessed tended not to quantify generally. Assessments can provide information of this kind only if holistic tasks such as practical investigation are used in conjunction with atomized, criterion-referenced tasks. This suggests that a range of different data collection methods is needed. In particular, we need case-study data to supplement and enhance our ability to interpret the more generalizable survey data.

The fear that 'content-independent' tasks are invalid has resulted in many assessment initiatives in the UK overemphasizing content. This has a two-fold effect. First, it fails to provide information about pupils' procedural understanding. Second, it fails to uncover the complexity of process-content interactions for the novice who is attempting to solve science problems. In general, content-led assessments can only represent the extent to which children hold 'expert,' explicit knowledge and communicate it in 'expert' ways. This leaves the achievements of the novice largely unrepresented in assessment outcomes and, furthermore, provides very few insights into how children learn.

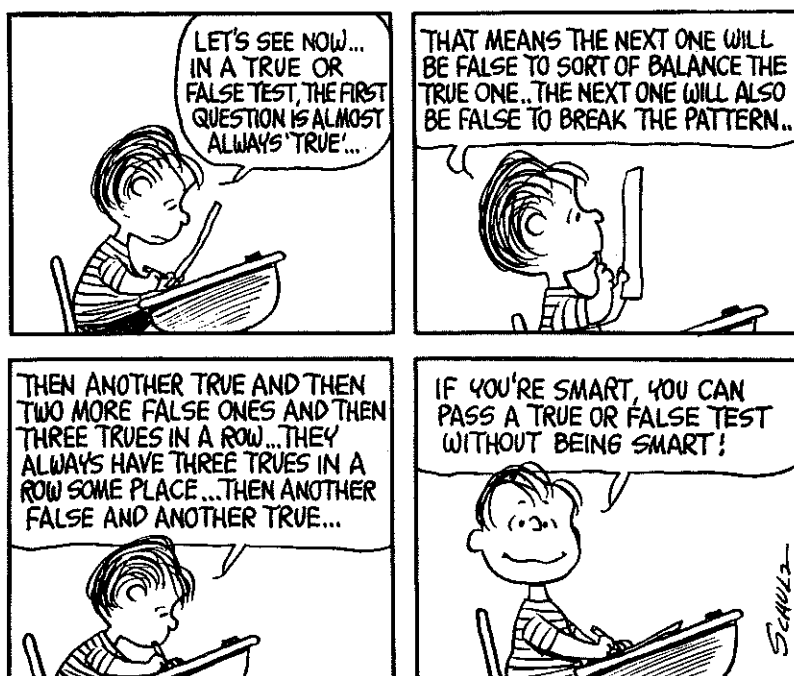
The results presented here have also touched briefly on the effects of using different cues in tasks and on alternative modes of presentation, operation, and response. Changes of this kind rather than representing alternative measures of the same criterion produce qualitatively different forms of data. In particular, the use of cues can alter the demands within any one assessment criterion

being tested and thus can raise or lower performance levels. On the other hand, alternative modes of presentation and response can shift the task from one test category to another. The surveys provide very strong evidence that we need a variety of data collection forms to represent pupils' achievements validly. Also, some forms of assessment provide deeper and broader insights into pupils' achievements, as the results in the Performing Investigations category revealed.

The survey results provide much detail about the factors that affect pupils' success in interpreting presented data or identifying the significant variables within investigations. One factor arises throughout the APU results: Pupils have problems understanding specific variables and the relationships between them in various investigations. The observation tasks discussed provide a variety of ways of establishing the type of understanding pupils hold. Open tasks that cover a range of content allow pupils to select what they consider noteworthy and to demonstrate how their perception of what is relevant varies as content and context change. Such tasks will not 'fit' into a unidimensional profile element or test category in science, yet the data they generate are necessary if we are to better interpret pupils' procedural errors in other assessment situations and to discern the alternative problems (to those of the assessor) that pupils commonly see and pursue. One of the main findings of the APU science surveys is that the lower achieving 40 percent of the population at each age do not generally fail by providing incorrect responses but by providing responses to alternative tasks. A failure to identify pupils' tasks not only results in invalid assessment practice but also means that pupils are provided with classroom learning situations that they cannot make sense of. If pupils cannot make sense of school activities, they do not have the opportunity to learn from them.

PART FOUR

Assessment in Science Education Research and Development



Introduction

George E. Hein

Those engaged in teaching and in administering schools are not the only professionals who need to develop methods for assessing children's knowledge and understanding of science. Curriculum developers and researchers also face this problem. It is illuminating to note how these professionals address this concern. Fortunately, there is currently considerable activity in these areas.

In 1987, the National Science Foundation (NSF) began a new initiative for developing elementary school science curriculum materials, bringing educators, curriculum developers, and publishers together in partnerships to provide a new generation of materials. At the time of this conference, three of the eight funded projects had been in existence long enough to have evaluation staff and to have some materials under field test. Staff from two of these three projects have contributed to this volume.

The extent of agreement among these projects is remarkable. Each has recognized that traditional measures of science achievement would be inadequate to describe the projects' outcomes. As they work with children in classrooms, each project's researchers use a variety of assessment means, including performance measures, open-ended written questions, interviews with children, observations in classrooms, and other approaches. In the next paper, Maryellen Harmon and Jan Mokros discuss all of these, as well as the reasons projects have adopted this multifaceted approach to evaluation.

The reasons the project evaluators provide for their choice of methods are also strikingly similar. Each recognizes the crucial role that the teacher plays in science pedagogy, and that these assessment methods will provide means for empowering teachers and validating their work. They also affirm that more in-depth descriptions of children's learning may have a positive political impact in promoting science education, and may assist the adoption of these projects in school systems across the country. Finally, they agree that we must employ a wide range of assessment methods if we are to assess the learning of the entire student population rather than that of a selected few.

Another area that provides information relevant to assessment is research on children's science concepts. In the last decade, several groups in many countries have begun careful studies of what children (and adults) know about the world, how their beliefs relate to current scientific views of the nature of the world, and the impact of instruction on these views. Rosalind Driver describes the current state of this research and emphasizes the significance of this work for assessment. Both the results of this research (which demonstrate that many naive views are deeply held, often differ from accepted scientific opinion, and are difficult to modify even with instruction), and the methods used to discover what children believe are important for any discussion of assessment issues. The preferred methods of researchers in this field clearly differ dramatically from the paper-and-pencil, short-answer tests that dominate science instruction.

Assessment in the New NSF Elementary Science Curricula: An Emerging Role

Maryellen Harmon and Jan Mokros

Today there is widespread discontent with the capacity of existing national norm-referenced tests to measure the kinds of scientific learning in the new, hands-on curricula in elementary schools. The goals of the new round of curriculum projects funded by the National Science Foundation (NSF) go beyond information to emphasize processes and problem solving, integration and meaning-making, and the understanding of a few basic concepts. Such goals are typically trivialized or ignored by multiple-choice tests and particularly by those geared to show success in a broad national population. Even the definition of success is different from the involvement, questioning, and problem solving that are indicators of success for those engaged in hands-on science. Clearly, an alternative to the existing testing models is needed.

However, an alternative has not yet appeared on the national scene. In this paper, we will discuss the emerging role of assessment in supporting the new wave of science curricula funded by the NSF and the success of some of the models for alternative modes of assessment.

Assessment has traditionally played a follower role. That is, after a piece of curriculum or a consensus about outcomes has been developed, evaluation experts create pretests and posttests based on that curriculum to measure the level of achievement of the desired outcomes. In many of the new generation of NSF-funded projects, the assessment team acts in advance of curriculum development to determine what needs, understandings, and misconceptions students bring with them, and therefore what direction the curriculum development needs to take. In addition, as Tyler reported some 40 years ago (Madaus and Stufflebeam, 1989), the evaluation expert works along with the writers, as a member of the core team, to help clarify the outcomes desired and the indicators for these outcomes, to elicit students' and teachers' misconceptions, and to explore the limits to freedom of activity and discourse acceptable in the particular science classroom. The information gained helps curriculum writers design activities carefully directed toward a chosen outcome, rather than a series of episodic, nondirected, but "interesting" experiences. At this stage of the writing,

Maryellen Harmon has been involved with science and mathematics education at the elementary, secondary, college, and graduate levels for 40 years. She is presently at the Center for the Study of Testing, Evaluation, and Educational Policy at Boston College and is working on two major studies of assessment funded by the National Science Foundation.

Janice Mokros is a developmental psychologist who codirects the Mathematics Center at Technical Education Research Centers. She was director of evaluation for the NSF-funded National Geographic Kids Network Project.

assessment suggestions for teachers are generated. These suggestions seed or reinforce the principle that assessment is not for judgment but for instructional decisionmaking: Today's assessment requires decisions that affect both the content and the pedagogy of tomorrow's instruction. To serve such a function, assessment cannot be confined to measuring student achievement on standardized tests. Instead, it must be tailored to the developmental stages, conditions, and multiple faces of the curriculum as it is taught in the classroom.

In broad terms, such tailoring includes assessment of:

- the preconditions of instruction
- the evolving curriculum
- the classroom effectiveness of that curriculum, as measured by both affective and cognitive tools.

By preconditions we mean the present student interpretations or "misconceptions" of phenomena, their experience of the world, their present "knowledge" of scientific facts, and the classroom culture that varies from school to school and determines the way science is done. Other preconditions for curriculum writers are revealed in teacher perceptions of the classroom, of their own situations, of pedagogy in general, and of "hands-on" science teaching in particular, as well as in the teacher's competence in hands-on science teaching and learning. Related concerns include teachers' need for support and/or inservice and the readiness of the district both to support activity-based science and to allow teachers the freedom to help shape a new approach to old problems.

It is worth reiterating that assessment theory requires a clear perception of the outcomes desired from the use of a particular piece of curriculum, strategy, or mode of instruction. To obtain both clarity and consensus as to the outcomes and the indicators of those outcomes is one of the most difficult tasks faced by the assessment—for one must steer between the Scylla of a mechanical, fact-oriented, and easily measured behavior, and the Charybdis of goals so broad and all-encompassing as to make progress toward them virtually unmeasurable.

The theory of assessment has clearly moved forward in the new NSF projects, and learning objectives, as formerly expressed, are no longer as useful as they were when curriculum developers were operating under behavioral paradigms. Today we need to go deeper, to separate the descriptive and prescriptive, to probe conceptions and misconceptions, and to create alternative modes of assessment so that broad outcomes can manifest themselves in a number of ways.

Although a stronger role is emerging for assessment in the formative stages of curriculum development, the assessment task has only just begun. Constructing instruments that reveal cognitive understanding—the presence, absence, or fuzziness of concepts—is not too difficult. Often, the most effective instruments for doing so are not pencil-and-paper tests but some other means of eliciting demonstration of skill and utilization/application of scientific processes—performance assessment. Here, elementary school education has lagged far behind the “practical” examinations required in all vocational schools, including those in the arts as well as in engineering, nursing, and other professions. Recitals, portfolios, and other performance assessments must demonstrate acquisition of knowledge, concepts, and skills, and the ability to apply them appropriately in “untaught” situations. The trouble with performance assessment in elementary schools lies in the constraints upon use of such methods, constraints created by teachers’ “dis-ease” with such instruments, their lack of time to incorporate the assessment process into the course of instruction (embedded assessment), and above all, their lack of experience with this mode and the ways of interpreting and using the results. We gain by introducing such a mode: When students must apply knowledge rather than regurgitate it, they not only demonstrate mastery of concepts but also call upon problem-solving skills—a clear objective of science education. However, a change in the way we do assessment will challenge our most creative insights concerning teacher development and the dynamics of school change.

Evidence shows that what is tested has a controlling influence over what is taught in school districts across the nation. That science is not taught, or is sacrificed in favor of other subjects (Weiss, 1977, 1987; Mullis, et al., 1988; Raizen, 1988), in a large percentage¹ of the nations’ elementary classrooms reflects, in part, the fact that it is not valued enough even to appear on the nationally normed standardized tests at certain grade levels. School districts using batteries that include science frequently report that they do not use the science portion of the battery. Even where science is tested (using tests published by Science Research Associates, Iowa Tests of Basic Skills, California Achievement tests, and others), research has shown that nationally normed achievement tests do not measure the broad range of processes, cognitive structuring, and higher order thinking skills that are the object of the newer science programs (Raizen, 1989; Hein, 1988; Schwartz, 1977; Taylor, 1977). On the contrary, by their emphasis on the types of questions that can be answered by simple recall of facts and recognition of experiments described in the textbook, these

tests may well militate against higher-order thinking and a less predictable hands-on approach in teaching. With the type of test presently in existence, students who have been drilled in a broad range of facts will do better than students whose more focused studies involve questioning, experimenting, and trying to interpret their own results: in short, learning to do science and to think critically. The existing national norm-referenced tests do not support or encourage these skills or support the implementation of new insights and development in science curriculum materials or pedagogy. Their continued, quasi-universal use may dampen, if not totally inhibit, implementation of newer approaches. Alternatives are needed not just for those school systems already committed to implementing new approaches, but also as a means of encouraging other districts to risk a more effective mode of science teaching.

This paper addresses the need for alternatives to existing norm-referenced tests, alternatives that will have equivalent or even higher quality and marketability than those presently dominating the market, but will be more congruent with the type of science teaching the National Science Foundation has promoted. As evaluators of NSF curriculum projects, we are working toward developing assessment strategies and materials that are:

- convenient for teachers to use but not limited to objective items;
- able to measure science process skills as well as content mastery;
- sensitive to the needs of diverse populations including children with special needs, limited reading proficiency, second language backgrounds, and other at-risk populations;
- multimodal: going beyond the traditional paper-pencil tests to include other formats such as short essays, extended problems, performance assessments, journals, charts, simulations, and software.

Research into alternative formats for assessment is not lacking, but the serious development of these research insights into practical testing programs capable of national dissemination and implementation has not been undertaken. The reasons are obvious. Developing alternative assessments requires time, resources, and intensive collaboration among test publishers, the testing research community, and science educators. It requires an effective educational and marketing strategy to bring new types of testing into common use by people in schools and on school boards across the

nation, people who are at present relatively unaware of the urgent need for change. Those who are aware, thanks to the sense of urgency created by the popular press, frequently address the needs with policies advocating more of the same: drill-and-practice approaches, “back to basics,” “eliminate frills” (including hands-on science?), and “improve standardized test scores” regardless of what they measure.

THE NSF ELEMENTARY SCIENCE CURRICULUM PROJECTS

The purpose of the latest generation of NSF-funded projects is to improve the quality of nationwide science education in grades K-6. As assessment people, we have an additional goal: to improve the quality of assessment in grades K-6 and thereby improve the quality of science teaching in all elementary schools. We are attempting to develop assessment programs that will be broad enough to encompass the following instructional characteristics, characteristics inherent in all our projects:

- Topics and activities must have personal meaning and social relevance for students. By this means students will be able to integrate their new concepts and thought processes into their own cognitive matrix and be able to apply them as needed in other domains of their daily lives and future studies.
- In order to help students become in-depth learners, we will present fewer topics, but cover them in greater detail. Students learn concepts most effectively when they experience them at many levels, using many senses. Therefore, a hands-on approach is essential. Because a hands-on approach requires more time, less time is available for the kind of broad coverage reflected in existing testing instruments.
- Science is integrated with other subjects, specifically language arts and mathematics.
- Decisions about content, processes, and skills are based on the developmental levels of the students. Development here is not limited to Piagetian operations but is understood to include social and ethical, as well as cognitive, development.
- Students have opportunities and assistance to explore natural phenomena, to question the data,

to solve problems, and to structure meaning from experience. No longer is it satisfactory to expect them merely to learn the descriptions and results of other peoples' thinking. They need to learn to think, process, question and integrate for themselves.

- Instruction builds on students' present understandings and/or misconceptions and aims at enabling the students to develop new explanations, attitudes, and skills.

In both of the projects reported upon below, a major thrust of the evaluations is to develop student assessment techniques that are both congruent with the new science pedagogy and feasible to implement in the classroom. The projects' assessment approaches are responsive to teachers' own needs for ongoing documentation of student learning. At the same time, the assessments being developed help teachers see the need to go beyond traditional testing formats and provide them with alternatives that yield a deeper understanding of what students are learning.

These pioneering, NSF-funded elementary curriculum projects are based on a three-way partnership between schools, developer, and publisher. From the assessment standpoint, such a partnership is extremely advantageous: It means that research insights will be translated into classroom practice, which in turn can be disseminated by the publishing community. The projects are providing a leadership role in linking what is taught and what is tested in elementary science classrooms.

IMPROVING URBAN ELEMENTARY SCIENCE: THE IUES PROJECT

Education Development Center (EDC), supported by a grant from NSF, has begun a project to develop a science curriculum for elementary schools nationwide. This new program is directed toward both developing scientific conceptual understanding and improving creative and critical thinking and problem-solving ability by experimental work in a natural environment.

The program has several unusual features. Its target is the urban system in cities challenged by large numbers of low-income, minority, and special-needs students. Its design is collaborative, with teachers and curriculum developers working together to design curriculum that works and that *will* be used. One of its driving principles is that curriculum writers must discover—and construct curriculum to address—children's naive theories or misconceptions about their world, and must do this in ways that

relate directly to the urban world. Another principle is that learning is holistic, thus science must be integrated with the rest of the elementary curriculum, particularly with language arts and mathematics; a third, that group learning and group assessment are to be facilitated wherever appropriate. Behind this last principle lies the conviction that children are as social as adults and learn constructively from, and with, one another—learning promoted by the affect of cooperation and support in the group.

The project involves teacher development teams from San Francisco, Montgomery County (Maryland), Boston, and Cleveland who are participating in the design and review of the new activity-based modules for grades K-6. Additional teams of teachers from Los Angeles, Baltimore, and Hartford will provide input and field-test feedback to ensure that the final curriculum meets the needs of a number of urban settings, operating under a variety of state and local mandates. These teachers have not been involved in the development of the project and have only the printed modules to work from. Hence, they provide data on the potential effectiveness of the curriculum in its published form, without intensive inservice training. The entire process is ongoing: at present, four modules have been field tested; others have been trialed in classrooms and have undergone major revisions before field testing; and four new modules are being developed in collaboration with classroom teachers.

Materials have also been submitted at scheduled intervals for review by a distinguished advisory panel of scientists, science educators, cognitive psychologists, education researchers, classroom teachers, and administrators.

Field tests for Modules 1-4 were conducted in the fall semester of 1988. Feedback underscores the challenge for writers. The dilemma: whether to bring curriculum and assessment down to the level of what teachers are presently accustomed to and expect children to be able to achieve in a typical urban classroom, or whether to use curriculum and assessment to elevate the level of expectations, challenge, and achievement. Although teachers' insight and experience are important in resolving this dilemma, following too closely the criticisms of teachers whose field of study has not been science or science education can lead to watered-down repetitions of existing and familiar programs.

A Lead Role for Assessment in the IUES Project

One of the ways in which assessment is playing a lead role in the development of IUES modules is in the use and interpretation of probes.

Prior to the completion of a module, the evaluators develop probes: short, open-ended pretests designed to elicit students' present interpretations of the concepts planned for the module. If misconceptions exist, they are identified at this point for guidance of the module developers. For instance, in response to a question about where a plant gets its food, some students may say "from the supermarket: My mother buys plant food," while others may say "from the soil" or "from the air." This information is given to the module writers so that they can address the issues in the teacher's guide.

Evaluators also work with writers to identify key concepts, their interrelatedness, necessary scientific information, and the scientific processes we wish to stress in each module. This collaboration enables the evaluators to develop embedded assessments; grids or checklists to facilitate teacher recordkeeping; daily assessment suggestions; and extended problems.²

After the module is written, the evaluator assumes a more traditional role by producing pre- and posttests, observing classrooms, and conducting student and teacher interviews with a view to determining both affective and cognitive outcomes of the module.

Pencil-and-paper instruments are used for pre- and posttests. Typically, pretests contain open-ended questions to permit students to reveal whatever concepts they now hold without being influenced by a selection-type question. Posttests comprise a number of modalities: a performance assessment, multiple-choice items, justified multiple-choice items,³ matching uneven columns, short essays, diagrams and tables, and concept maps. We have developed scoring schemes for each module to ensure comparability between essay answers. Replicability averages 0.85 across modules. Final performance assessments are structured: Students move from one prearranged station to another, performing and explaining certain tasks, such as wiring an electrical circuit in parallel or calibrating a measuring instrument.

Results to Date

The first purpose of all assessments at this stage is formative—to assist those who are rewriting the modules for further testing or for publication. An additional formative outcome is to enable the teacher to adjust pedagogy and timing to the students' levels of understanding. A second purpose is to enable the evaluators:

- to collect data on the usefulness and feasibility of alternative modes of assessment, such as embedded assessments, structured performance assessments, simulations, and concept maps;
- to validate, by interview and observations, the information obtained on written instruments;
- to revise and refine the instruments.

The work is still in process: eight of the proposed modules have not yet been written; two modules, "Circuits and Pathways" and "Plants," have been completed, field-tested, and analyzed; another two have been field-tested; four more have completed their first piloting; and four are presently in classrooms for the first time.

In general, the hands-on approach embodied in these modules works well. Across all classes, the gains from pre- to posttests averaged 26 percentage points. Individual students gained as much as 40 percentage points. Forty percent of the students who had shown level 2 responses (misconceptions) on the pretests shifted to level 4 (a complete and accurate response) on the posttest. However, certain misconceptions proved to be quite robust. For example:

- the belief that electric current is "used up in the bulb" and that there is no current beyond that point (22 percent of the respondents);
- the belief that the skin of the lima bean "is poisonous and should be peeled off" (16 percent);
- the belief that electric current flows in a circuit even when the switch is open (50 percent of the respondents).

An interesting phenomenon emerged. Most students who appeared quite ignorant of a topic on the pretest (a partial or slightly inaccurate response) moved to a level 3-4 response on the posttest. However, a small percentage (approximately 10 percent on each concept) expressed one of the classic misconceptions on the posttest. It seemed as though they had to go through this stage in their journey toward understanding. (Reminiscent of the historical movements of thought from a Ptolemaic to a Copernican universe or from a flat-earth theory to our present conception!)

Assessment data gleaned from the first four modules reveal a significant success in teaching the intended concepts. On one module, the range of increase in class averages between pre- and posttest was 27-42 percent among the various classes, with an average increase per class of 33 percent. On another, the range of increase across seven classes (175 students) was from 3 percent to

64 percent, with the average increase across all classes of 19 percent. Equally dramatic increases were measured for individual students.

The mode of scoring (a modification of the scheme used by the National Assessment of Educational Progress) is loaded to identify the presence of alternative concepts or “misconceptions,” as well as partial and complete responses, and enables us to identify confusions evident before and after instruction so that teaching materials can be rewritten.

Specific analyses have been made on a module-by-module basis, at both the student and the whole-class level. Teachers now receive the analysis of the pretest before beginning instruction so that they are alert to areas of student misconceptions.

A second purpose of assessment in this pilot/field test phase of the project is to collect data on the feasibility and usefulness of various modes of assessment. We have found that:

- At least initially, teachers are most comfortable with a multiple-choice format.
- Teachers in their second year with the project are comfortable with open-ended questions and alternative modes of assessment such as extended problems.
- Second-year teachers appreciate the value of the probes; first-year teachers are usually upset by their use because we are “testing something we have not taught yet.”
- Uneven matching columns, particularly those matching examples with definitions, were upsetting for the teachers (who wanted to, and sometimes did, rewrite the questions in ways that trivialized the content and matched the column length).
- Uneven matching column questions yielded important information about concepts that were fuzzy or not well-discriminated from related concepts in the students’ minds: i.e., series vs. parallel circuits; closed circuits vs. conductors. However, such questions were hard for the students—perhaps because students are not being pushed in class to look at what a concept does not mean as well as what it does.
- For a variety of reasons, chiefly time and readily available equipment/supplies, teachers avoid the

performance assessments AS ASSESSMENTS (skipped, data not returned). Oral responses indicate that they do not consider them real assessments.

- Teachers and students enjoyed, and teachers used with profit, the embedded assessments.
- Most teachers did not see the value or “did not have time” to maintain systematic records of growth in particular concepts and processes for individual students, whether by means of check-lists, grids, or personal notes. They claim “I can tell who is getting it.”
- A significant number of students had difficulty with questions involving tables of data. At least an equal number could read and interpret such data.

Data are still insufficient to comment on effectiveness and comparability of concept maps, drawings, simulations, and portfolios.

Since our intent is to empower teachers, all assessments are being conducted by the classroom teacher with the assistance, as needed, of the outside evaluator. The latter also collects data independently for comparison purposes and supplements written data with interviews.

In addition, the project assessment team plans to gather data not only at the end of the module but also at the end of the year, using both curriculum-specific and standardized tests. The collection of data on standardized tests seems to us necessary to determine if any correlation exists between project-specific test results and standardized test results. Since the content taught in the project will be different from most existing textbook curricula, we anticipate that much of the content tested on typical standardized tests will be irrelevant and that the coverage of these tests will be broader but more superficial than on the project-specific tests. However, knowing just what the comparative data are and, if possible, precisely why there are discrepancies is important. This information will also be useful to school district decision makers trying to make informed choices about what they want to happen in their science classrooms and how to explain these happenings to their publics.

All the above instruments and testing procedures aim at measuring curriculum effectiveness through student outcomes. It is also important to measure curriculum impact on school districts and schools in terms of manageability, organizational impact, and

costs in dollars, time, and other resources. This data will be collected by questionnaire, observation, and interviews with administrators as the project completes its field-test years. As indicated, comparative data on the effectiveness of various modes of teacher inservice, while not a specific part of the project, is planned as part of another research project.

Summary

The first goal of the Evaluation Plan for the IUES Project is to use assessment results to guide the shaping of the new curriculum into the most effective instrument possible for developing science concepts, attitudes, and skills within the constraints of existing classroom reality. This shaping process is integrated with, and sometimes leads, curriculum development.

The second goal is to develop assessment instruments that will provide an alternative to existing portions of standardized batteries, and which will measure a broad range of scientific processes and higher-order thinking skills, as well as conceptual knowledge.

The third goal goes beyond curriculum development to address a constraint that could prevent the use of the IUES curriculum. Testing is a reality of considerable political significance. Test results determine which schools get extra resources, what school improvement funds states allocate, how teachers and parents perceive the potential of individual children and cultivate or neglect that potential, how the media and public compare schools, and even (although usually not overtly) whether teachers or administrators are retained, praised, or transferred. If new modes of assessment can provide policy makers with more valid data than presently exist in order to choose curricula and justify educational decisions, it may help all of us obtain the necessary resources for improving effective science education.

Rarely do existing tests justify the uses made of the results. Nor, perhaps, are they designed to do so. In the interest of time, efficiency, and management, they force all students into a single (usually multiple-choice) expressive mode and assume that what cannot be demonstrated in this mode is somehow implied thereby or is not important to measure. On the other hand, many teachers are anti-assessment—perhaps because of fear that the results will be used to judge them and not the curriculum. There is a significant difference between allowing or providing alternatives in assessment and collapsing into a nonmeasurement stance that relies on “feel” and “class interest” as indicators of understanding.

Up to the present, most alternative modes of testing have been project-specific and confined to the particular researcher's area of interest. This is why we believe it important to explore, in depth, the relationship between the results of curriculum-specific tests in this project and nationally standardized tests. Until national tests are developed that assess different domains and allow for multiple, individual, and group modes of response, assessment in science will not be an instrument promoting scientific literacy nor will it be a means toward informed instructional decision making.

THE NATIONAL GEOGRAPHIC KIDS NETWORK PROJECT

The National Geographic Kids Network Project, developed by Technical Education Research Centers (TERC) and published by the National Geographic Society, is an exciting series of cooperative science experiments in which fourth through sixth graders use mapping and graphing software, along with a telecommunications network, to share and discuss data. The project is based on the premise that students should get first-hand experience with real and engaging scientific problems that have an important social context. Our goal is to convince students that they can and should be scientific thinkers: They are encouraged to collect, analyze, and discuss data. Using the telecommunications network, children learn that science is a cooperative venture in which they can participate.

The Kids Network project focuses on the development of a neglected set of "basic skills" in science—skills that allow children access to science and foster true scientific literacy. What we mean by basic skills is the essential skills that scientists use regularly in their everyday work. The skills we hope students will develop, and which we are currently assessing, encompass a broad range of scientific process skills. In the current units on acid rain and weather, we are particularly interested in helping students learn to:

- collect data in an accurate manner, making appropriate and reliable measurements;
- organize and represent information so that it makes the most sense to self and others;
- use maps, graphs, tables, and other displays to find patterns in their data;
- get a good "feel" for the data by asking questions about it and seeing it from different perspectives;
- discuss and write about observations and other data.

Our goal is to help students develop a new set of basic skills in these areas and to spark their interest and involvement in science so they will continue to develop as scientists.

Assessing the New "Basic Skills" in Science

How do we document and assess the extent to which students who participate in the Kids Network Project become more capable scientists? The task seems as complex as assessing the products and professional development of "real" scientists. To begin, it is imperative that we make our assessment congruent with our philosophy. As Valencia and Pearson recently pointed out, as long as assessment is based on one view and learning and instruction based on another, "we will nurture tension and confusion among those charged with the dual responsibility of instructional improvement and monitoring student achievement" (Valencia and Pearson, 1987). To get a flavor for the striking contradictions between our views about science learning and the traditional ways that many evaluators have assessed science learning (table 1). We must narrow the gap between what we are trying to teach and what we are trying to measure.

TABLE 1. Comparison of New View of Science Education
Versus Assessment Practices

<u>Views of Science Education</u>	<u>Limitations of Assessment Practices</u>
Children bring to the classroom a developed set of conceptions and misconceptions about science.	Yet, our assessments do not attempt to describe children's existing beliefs and knowledge, or how these change as a result of science instruction.
The ability to ask good questions of the data is essential.	Yet, assessments never give students the opportunity to develop new questions based on particular findings.
Doing science involves the integration of many skills (logic, knowledge of content, question-asking, finding patterns, using evidence, etc.).	Yet, assessments typically isolate each skill, never asking students to do science the way it would actually be done.
Doing science means being able to flexibly apply principles, theories, and existing knowledge in new situations.	Yet, assessments are "context bound." Students are asked to apply principles only in the familiar situations in which they were learned.
Doing science means being able to construct appropriate tests of real questions.	Yet, assessments typically are limited to asking students to identify the relevant variables in a prespecified experiment.
In science, there are often different (even conflicting) explanations for a particular phenomenon.	Yet, assessments almost universally demand one "correct" explanation from students.
Doing science means communicating about your findings, discussing different interpretations, and arguing about the usefulness and elegance of those interpretations.	Yet, assessments never invite children to address an audience other than the tester.

Those of us involved in the National Science Foundation projects are fortunate to be working with innovative curriculum developers who know how to get children truly involved in science. On the Kids Network project, we are also working with scientists who serve as models for children by demonstrating both the creative and the analytic aspects of science. In keeping with the spirit of the project, it is essential for us to develop assessments that capture both the creative and the analytic sides of doing science. To simply distill the project's goals into small chunks of measurable learning objectives is not enough. This process might accurately reflect some specific goals of the project, such as finding out whether children can use pH paper and compasses, and even whether they know the relation between wind direction and temperature change. But the bigger questions need to be asked. We want to know how children have progressed in their understanding of what it means to do science; whether they can take a question, gather data concerning this question, interpret the data, and discuss its meaning and limitations.

In planning the assessment for the Kids Network project, we have found a few science assessments that serve as excellent models. Most notably are the assessments developed by the APU and the NAEP groups. What these assessments have accomplished that others have not is noteworthy and merits some discussion. First, the items on these assessments are familiar, concrete, and compelling to children. Working on the problems in these assessments is not a boring exercise. For example, rather than asking about principles involved in heat required to change state, Driver's group (1974) asks about why it takes frozen peas longer to come to a boil than fresh peas. There is a similarly compelling item posed by a group of Swedish educators: Why does it take *longer* to cook potatoes in an oven *at a higher temperature* (375 degrees Fahrenheit) than it does to boil them in a saucepan *at a lower temperature* (e.g., the boiling point of 212 degrees Fahrenheit)? This is a counterintuitive yet very familiar problem, one that demands thinking rather than a memorized response. Unfortunately, most tests that supposedly tap "higher-order skills" ask children simply to discuss the merits of one theory versus another or to give an example of a particular principle. All of this can be done without thinking by the conscientious student who studies his "lines" ahead of time. But good problems demand on-the-spot thinking and application of a principle to a particular situation.

What is most impressive about tests such as those developed by the APU group is the care that has gone into analyzing the responses. Analyzing responses is perhaps the most difficult

problem for the creative assessor: It seems the more creative the item, the harder it is to score. Only in a few cases in science have we seen the careful "holistic scoring" that is much more widespread in writing assessment. Developing analysis strategies that are meaningful and rigorous is a demanding and expensive task.

In developing assessment tools for the Kids Network project, we are making use of the excellent models provided by the British projects. Specifically, our assessment work to date has focused on:

- developing post-unit tests that pose a limited number of compelling "story" problems that demand thinking and pattern-finding;
- interviewing children (pre- and postproject) to document changes in their perceptions about science, how science is done, and the role of science in solving human problems;
- reviewing children's real writing on the telecommunications network to get an idea of how they are thinking about their data-collection efforts.

Below is a brief description of our assessment techniques.

Story Problems: Finding Patterns

For the Kids Network unit on Acid Rain, we developed problems to determine children's level of sophistication with respect to understanding the causes and effects of acid rain. One item capitalizes on the "mystery substance" idea and asks students to analyze a local water sample and tell about its probable effects on seeds and on pennies. Another asks students to examine data on a map that shows wind patterns and emission sources and to make predictions about where acid rain will be the most serious problem. These problems were very closely linked with learning objectives of the unit. We also implored teachers to administer this "test" in a way that was suitable to the skills of their students. That is, we asked them to work individually with children who have difficulty in reading and writing so that these "mechanical" barriers would not interfere with our assessment of scientific understanding.

Forty teachers administered the test and sent us not only the student data but also their own evaluations of the tests and documentation of how they used it. Overall, teachers found the test quite useful, indicating it was similar to tests that they would construct to evaluate student learning. The teachers who needed to arrive at grades for the unit had an easy time arriving at scores on the test. Interestingly and disappointingly, the scoring schemes

developed by some teachers revealed that the teachers themselves did not understand some of the concepts to be taught in the unit. This in itself is important (albeit serendipitous) evaluation feedback that will be used in developing our teacher-training materials. The test has great value to us in demonstrating what students had learned and was clearly useful to the teachers as well.

The major problem with this assessment technique is that the costs for evaluating student responses are prohibitive. We have not had the time or resources to develop a rigorous framework for analysis. Ideally, we would want to review a sample of the tests, develop an analysis strategy, try out and refine this strategy on a bigger sample of student tests, revise the analysis strategy, and then share it with teachers. The reality is that we can only begin to develop an analysis plan, which can in turn be shared with and elaborated upon by teachers.

Interviews: What Does It Mean To Do Science?

Doing science means that one has to think like a scientist, understand how the scientific enterprise works, and value the significance of scientific contributions. A goal of our evaluation is to determine whether students have a better understanding of what it means to do science. One could approach this task by administering any one of the number of attitude assessments in the science education literature. However, most of these surveys are superficial or ask questions that have little to do with the scientific enterprise itself. (For example, one such instrument contains a multiple-choice question asking how a scientist might decide whether or not to go to a new movie!) We do not need ratings of the value of science to self and others—these ratings tend to reflect global feelings of like or dislike of science, or perhaps liking for a particular science class and teacher. We decided that to really understand what kids understand about the nature of science, we would have to interview them. Our work was made easier because we were able to use a modification of an interview constructed by Susan Carey and her colleagues.

The interviews were quite useful in describing what children came to understand about the process of science. We interviewed 16 children from local evaluation sites, both before and after their participation in the project, asking them a series of open-ended questions about the nature of science, what scientists do, and what methods they employ. Several interesting patterns emerged from the preliminary analysis. We found that students began to have more accurate perceptions of scientific experimentation and were less likely by the end of the project to think of science as

mindless mixing and pouring. We also found that students came to believe that the work of scientists is important in solving human problems. At the posttest interview, they gave more specific examples of how scientists might address human problems. One example, from a fourth grader, illustrates this change nicely:

Preinterview: "What do scientists study?" Mark:
"... they study things that people aren't really
interested in, like gravity."

Postinterview: Mark: "Scientists figure out things
that are problems around the world. If there is, for
instance, the San Andreas fault they may want to
find out what is happening or how to fix it. If
someone is having a problem and they needed
help and they couldn't find anyone to do it they
would ask a scientist to study it."

A final intriguing finding came in response to the question "Do you know any scientists?" Although most students at both interviews said that they did not know any scientists, a few had changed their minds at the final interview and said that either they, their classmates, or their teachers were scientists.

We are continuing to use this interview in conjunction with our unit "Investigate," in which students identify a scientific problem they wish to investigate, collect data regarding this problem (using peers on the network as fellow data-gatherers), and analyze their findings. In this unit, students are operating very much like scientists, with a problem of their own choosing. Because participating students are in nonlocal sites, we are conducting the "interviews" via the telecommunications network, with students responding to and sending the questions via electronic mail. This method of assessment dovetails nicely with the writing assessment approach described below.

Writing Samples

A final method we are employing to evaluate student learning is an examination of the telecommunications record to determine how students write about their data and experiments. Our initial work, which examined students' spontaneous writing on the network, posed problems. Students vary greatly in how they use the network. This variability is attributable to resources, access, and the nature of teacher-given assignments. In order to use the network writing more effectively as an assessment technique, we found that we needed to build in specific writing assignments (embedded assessment) at particular places in the curriculum. We

have taken this more structured approach, which promises to yield a better picture of what students are learning. But even with this more structured approach, we need to develop better implementation strategies for teachers to ensure that children get a chance to use the network. For example, we have found that some teachers do not allow children to send letters to each other until the grammar and typing are totally correct. Needless to say, these specifications interfere with the amount of writing that children contribute and with the quality and depth of their discussions.

In conclusion, while we began with the three assessment approaches described above, our assessment approaches are evolving with the curriculum. The degree of emphasis on specific assessment techniques varies from unit to unit. And it is becoming increasingly clear in the Kids Network project that the six-to-eight week curriculum "unit" should be our evaluation focus. The idea of one, grand summative evaluation is not particularly useful in assessing student learning. Students' interest and understanding vary for each unit, and involvement with several units may not have an additive effect. Understanding of what science means may "click" for one student as she collects acid rain data, for another as he compares data on the amount of garbage that is generated in the lunchrooms of various schools, and for a third as she uses data she has collected to make predictions about tomorrow's weather. We can best identify the kinds of learning experiences that lead to significant change if we capture the transformations in students' understanding as close as possible to the time that these transformations occur.

CONCLUSIONS

Because the NSF science projects share a common spirit of inquiry, it is fitting that the projects' assessments have a common focus on student inquiry and problem-solving. Students are encouraged to think as scientists, and project researchers attempt to capture this scientific thinking and doing in all of the assessments. Our curricula, as well as our assessments, are heavily based on hands-on activities that allow students to demonstrate how they are thinking—rather than show us if they have memorized a prescribed scientific method.

Another theme throughout the evaluations is the need to make assessment data useful to teachers as well as to other researchers and project staff. All three projects have carefully designed assessment tasks that can be used and interpreted by teachers as they carry out their regular classroom work. In some cases, this has meant designing assessment activities that are

“embedded” in the curriculum. In other instances, students’ writing, lab work, or use of materials is documented for later analysis. Regardless of the technique, it is noteworthy that both projects have made the curriculum and its activities as congruent as possible with the assessment. In doing so, they have made the teacher’s role in assessment more feasible, and have at the same time created more meaningful and engaging assessment problems for students.

Assessment is indeed taking on a new role in these projects: It is a role that is more integrated with instruction, with the realities of the classroom, and, most importantly, with the developmental needs of students. And the assessments are more integrated with the nature of science: Science is a discipline that demands creativity, flexibility, and clear thinking—not simply memorization, patience, and adherence to prescribed procedures. As evaluators, we must design assessments that reflect the reality of what it means to do science.

Assessing the Progress of Children's Understanding in Science: A Developmental Perspective

Rosalind Driver

INTRODUCTION

An extensive literature built up in recent years indicates that children develop ideas about natural phenomena before they are taught science in school (Pfundt and Duit, 1985; Jung, et al., 1982; Helm and Novak, 1983; Driver and Erickson, 1983; Gilbert and Watts, 1983; Driver, et al., 1985). In some instances, these notions (variously described as preconceptions, misconceptions, intuitions, alternative conceptions, alternative frameworks, naive theories, or spontaneous reasoning) are in keeping with accepted scientific ideas. In many cases, however, children's notions and school science differ significantly.

Surveys undertaken in various countries have identified commonalities in children's ideas, and developmental studies are giving insights into the characteristic ways in which these ideas progress during the childhood years (Carey, 1985; Strauss and Stavy, 1982). In-depth investigations indicate that such ideas are more than misinformation: Children construe phenomena in ways that differ substantially from school science. They may continue to do so into adulthood despite formal teaching.

Currently, research on children's ideas is being interpreted within a cognitive science perspective (Carey, 1986). Key to this interpretation is the notion that human beings interpret situations in their world (whether text, dialogue, or events) in terms of *mental representations*. Moreover, as Bereiter (1985) suggests, "a core belief in contemporary approaches to learning is that knowledge and cognitive strategies are actively constructed by the learner" (p 201). Learning is seen as an active process whereby the learner relates existing 'mental representations' to new situations in order to construct meaning. The meaning so constructed depends on both the situation and the 'representations' the learner has available. Thus, from an educational point of view, understanding children's ideas prior to teaching is important because these ideas influence subsequent learning.

Furthermore, studies of children's reasoning about natural phenomena suggest that these mental representations tend to be specific to particular content domains. As Rumelhart and Norman (1981) argue: "Our ability to reason and use our knowledge

Rosalind Driver is Professor of Science Education and Director of the Children's Learning in Science Project at the Centre for Studies in Science and Mathematics Education, University of Leeds, UK. She has studied the ideas that children bring to their science lessons and the ways children's science conceptions develop during schooling.

appears to depend strongly on the context in which the knowledge is required. Most of the reasoning we do apparently does not involve the application of general purpose reasoning skills. Rather it seems that most of our reasoning ability is tied to particular bodies of knowledge”(p 338).

This perspective on the development of children’s reasoning in science suggests that progress in competence comes through the development of knowledge schemes within particular domains rather than in the development of general reasoning skills (as, for example, described by Piagetian stages). This has implications for both teaching and assessment.

This paper gives an overview of the research on children’s ideas about natural phenomena in a number of domains. It reports on one well researched area, that of force and motion, in greater detail than others in order to illustrate the range of data collection methods being used. It also outlines a number of general characteristics of children’s ideas about natural phenomena and, finally, identifies some issues of pedagogy assessment.

SPONTANEOUS REASONING ABOUT FORCE AND MOTION

A number of features characterize spontaneous reasoning in this domain.

The Force of Moving Objects

Researchers have found an intuitive association between force and motion to be very pervasive in the thinking of both children and adults. People tend to associate a force with any moving object. They see this force as what keeps the object moving and believe that it may get used up as an object such as a freewheeling bicycle slows down and stops. An 11-year-old clearly demonstrated this belief when asked, “What makes a ball rolling along the floor eventually stop?” He answered: “I don’t know—why do they stop—it’s just they always stop. After you push it they go as far as the push—how hard it was—and after that wears off it just goes back like it used to be.”

In probing the pervasiveness of this notion, Watts and Zylbersztajn (1981) surveyed 14-year-old English secondary school students’ ideas about force and motion. The pupils were given a number of questions relating to the forces on a cannon ball in flight (figure 1). In most cases (about 85 percent), students selected answers in which the force was associated with the direction of motion of the cannon ball.

In a seminal study in this area, Viennot (1979) presented French, Belgian, and British secondary and university students

with a number of written questions concerning motion. In one of the questions, shown in figure 2, six juggler's balls are drawn at the same height above the ground but at different points on their trajectories. A common feature in the students' answers was that the forces on the balls would be different because their velocities are different.

The converse of the rule that motion implies a force is, of course, *that there will be no motion if there is no force*. McCloskey (1983) reports a series of investigations in which he and his collaborators probed students' "knowledge-in-action." They asked university physics and nonphysics students to release a ball while moving across the floor so that the ball hits a target marked on the floor (figure 3). Researchers noted the number of students releasing the ball before, over, and after reaching the target. The results indicated that the majority of students released the ball directly over the target, suggesting that they may be neglecting the horizontal component of the motion of the ball or implicitly assuming it will be zero as soon as it leaves their hand.

A further feature of spontaneous reasoning in this area is that students believe that objects go in the direction they are pushed. Di Sessa (1982) reports on a study of students' interactions with a computer program called Dynaturtle. An object on a screen, the dynaturtle, obeys Newtonian laws of motion in that it remains at rest or moves in a straight line when no force is acting on it. It can, however, be given a 'kick' of varying magnitude and duration. When asked to move the dynaturtle on the screen so as to strike a target, students commonly ignore the initial motion of the turtle and direct the kick straight at the target (the expectation being that the turtle will move in the direction of the kick).

Objects at Rest

The association between the spontaneous ideas of force and motion emerges when students are asked to consider objects at rest on a surface. In discussing the case of a book resting on a table, for example, students acknowledge the existence of a downward force due to the weight of the book. ("If the table were not there, the book would fall down.") However, they tend not to identify the table as exerting an upward force on the book ("How can it if it can't move?") (Clement, 1983; Minstrell, 1982).

Weight, Gravity, and Trajectories

Researchers have also explored students' notions about weight and gravity. Students view heavier things as falling faster than lighter things (Gunstone and White, 1981; Watts, 1982). Furthermore, gravity and falling are often associated with air and

Figure 1. Task on Projectiles Used by Watts and Zylbersztajn

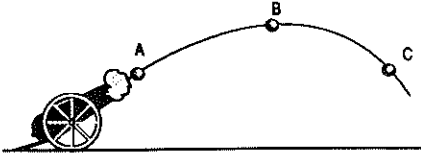
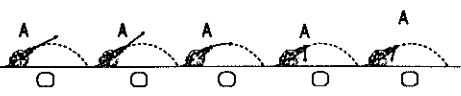
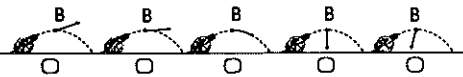
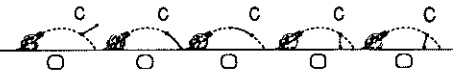
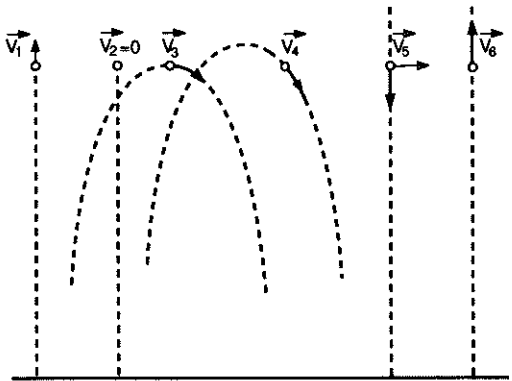
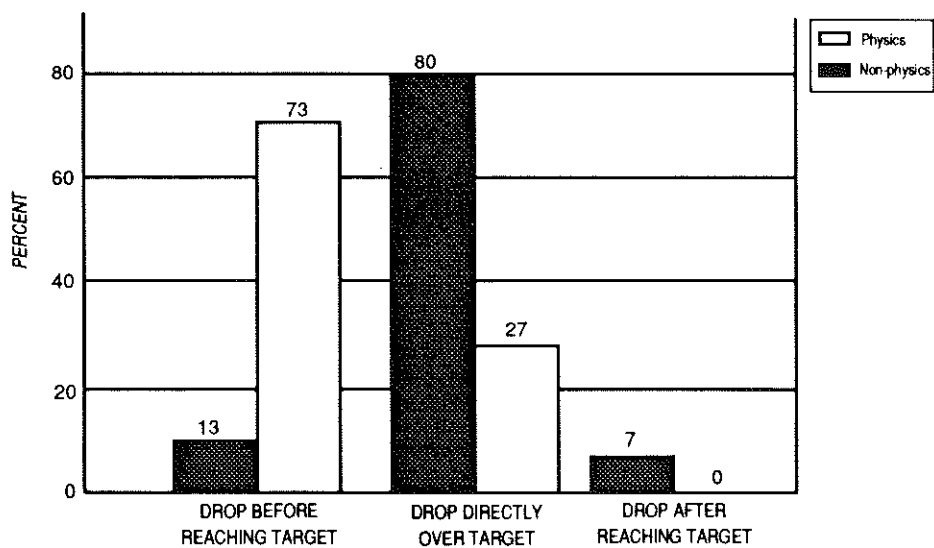
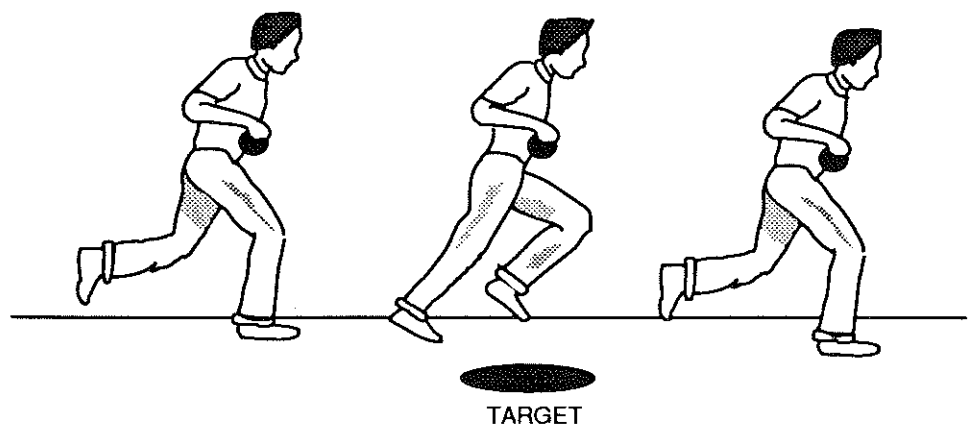
 <p>•A cannon ball is fired from a cannon. Points A, B and C are three different positions on the path of the ball. The following three cards refer to the situation.</p>	<p>•The arrows in the the picture are supposed to show the <u>direction of the force</u> on the cannon ball. Which picture do you think best shows the force on the ball as it is passing through point A?</p> <p>No Force</p>  <p>Explanation:</p>
<p>•Now, which picture do you think best shows the force on the ball as it is passing through point B (Its highest point)?</p> <p>No Force</p>  <p>Explanation:</p>	<p>•Now which picture do you think best shows the force on the ball as it is passing through point C?</p> <p>No Force</p>  <p>Explanation:</p>

FIGURE 2. Responses Given to Task of Predicting the Forces on Balls in Flight (Viennot 1979)



NUMBER OF STUDENTS RESPONDING	QUESTION RELATING TO	STUDENT'S YEAR OF STUDY	THE FORCES ARE...		
			equal	not equal	(no reply)
29	figure 1(p)	Last year of secondary school	39%	55%	6%
36		First year university	58%	42%	0%
226		First year university (Belgian)	44%	54%	2%

FIGURE 3. Predicting the Path of Falling Objects (Mcloskey 1983)



the atmosphere. When asked to predict the path of a projectile thrown in a vacuum, many Norwegian school and university students predicted that the path would be a straight line because “gravity needs a medium” (Sjoberg and Lie, 1981).

Researchers have identified the pervasive notion that “motion implies a force” in the responses of school and university physics students to a range of mechanics problems, including projectile motion and circular motion. Some have also drawn parallels between students’ ideas and ideas such as ‘impetus theory’ in the history of science (McCloskey, 1983). However, such parallels should be interpreted with caution.

SPONTANEOUS REASONING IN OTHER DOMAINS OF EXPERIENCE

Matter and Substance

Children’s ideas about matter and substance have been investigated from a number of perspectives. When a substance undergoes a simple transformation, such as when sugar dissolves in water, young children tend to think that the sugar disappears. Later they acknowledge that the sugar is still there even though it cannot be seen; however, they may think that it is weightless and does not occupy space (Holding, 1987). Young children also often believe that a substance disappears when it burns or corrodes. (“It burns up—leaving only ash—a part which does not burn.”) Older children, however, construe the continued existence of matter even when they cannot perceive it directly. They also begin to consider matter as being composed of discrete “bits” that can be dispersed and brought together again. However, children tend to see these bits as having the characteristic properties of the substance itself (e.g., they can expand when heated, melted, or burned) and do not therefore hold a scientific atomistic view (Brook and Driver, 1984).

Light and Sight

In the domain of light and vision (Guesne, 1985), young children see light as only a source (an electric light bulb, the sun) or an effect (a bright patch on the wall). They do not consider light as existing in space or traveling out from the source. Children first construe light as traveling when they consider luminous objects. They think of these as giving out light, but believe that the light can travel only a certain distance before it loses its strength, and further, that it travels further at night, when it is dark, than in the daytime. The connection children make between light and sight is indirect; they sometimes use notions of visual rays from the eye to an illuminated object to explain vision.

Heat and Temperature

Children tend to reason about phenomena in which objects or substances are heated in terms of heat as a quasi-substance (Erickson, 1979, 1980) that "flows through," "spreads out" and "fills up" objects. They may even see "hot" and "cold" as distinct and view temperature as a property of different substances. For example, metals may be identified as naturally colder than other materials such as wood or plastic.

Strauss and Stavy (1981) undertook an interesting series of investigations into the development of children's understanding of temperature. They document a U-shaped developmental pattern as children aged 4-13 differentiate the notion of temperature as an intrinsic property of substance from the amount of substance present. Parallels have been drawn in this area, as with mechanics, between children's spontaneous reasoning and historical developments in this field of science (Wiser and Carey, 1983).

Air and Air Pressure

Children between the ages of four and five identify air when it is moving, when there is wind or a draft. They do not tend to recognize the all-pervasiveness of static air. By age eight or nine, children begin to recognize that air exists in static situations and that it takes up space. For example, water will not enter an inverted bottle when lowered into a bowl of water because of the air that is already in the bottle; younger children find this incomprehensible and even argue that they see water entering. At the age of eight or nine, children also appreciate that air can be squashed (as in a bicycle pump with a blocked outlet). Although children at this age acknowledge that air has substance and takes up space, they see it as having either no weight or negative weight (air makes things lighter, as when air is trapped in floating objects.) The notion of weight used by children at this age is that of a solid object that can fall down or push down on a surface. This notion of air as weightless tends to persist for most children through schooling. A minority of students, however, do reconstruct their concept of weight and hence identify air as having weight by the age of 15 to 16 years (Brook and Driver, 1989; Séré, 1985).

Researchers have investigated spontaneous reasoning in a range of other areas, including electric circuits (Shipstone, 1985), the earth in space (Nussbaum, 1985), energy (Solomon, 1982; Watts, 1983), plant nutrition (Bell, 1985), and inheritance (Engel-Clough and Wood-Robinson, 1985).

GENERAL FEATURES OF STUDENTS' CONCEPTIONS IN SCIENCE

As this research suggests, humans do develop conceptions about a range of natural phenomena independently of formal instruction. Researchers have noted similarities in the conceptions used by children in different countries and from different social backgrounds prompting speculation about origins. Some have suggested regularities in children's experiences with physical phenomena as a contributory factor.

Conceptions pupils use to make predictions and explain events also appear to be influenced by various contextual features. Although the extent to which pupils use consistent models varies across domains of experience, they generally use quite different ideas to explain situations that are similar from a scientific point of view (Engel-Clough and Driver, 1986). In open problem-solving situations in classrooms, pupils often draw on and try out a range of possible ways of modeling a situation, each of which they check against the evidence for its "fit." In these practical situations, children are thus drawing on and checking models in a dynamic way. It is a matter of current interest how particular features in presented situations cue particular choices among models available to pupils.

Although they may differ from currently accepted scientific ideas, children's conceptions are coherent within a limited range of experiences and, in this way, "make sense." In the area of mechanics, for example, the notion that the force in a moving object gets used up is well-adapted to a world with friction. Recognizing the extent to which children's ideas fit their experience has important implications for educators because children may not appreciate the need to change models that seem to work effectively. This may account for the extent to which ideas in certain domains persist despite instruction and may explain why undergraduate science students still use certain 'spontaneous' notions in solving mechanics or electrical problems.

The notion, suggested by Solomon (1983), that pupils may bring different conceptions to interpret 'life-world' and 'school' science may also account for the persistence of naive conceptions.

However, some naive conceptions change as children get older, and studies of pupils' ideas in selected domains during the school years indicate how. A dominant perspective is that these changes involve radical restructurings of pupils' conceptions (Carey, 1985). The changes in pupils' ideas about light described earlier are an example of how children construe progressively more complex entities to account for perceived phenomena. Young children seem to have no notion of light existing in space; as they

get older, they expand their conceptions to incorporate first the notion of a "bath" of light and then the notion of light as "traveling."

Concepts of matter undergo a similar evolution. As children get older, they use the notion of invisible "bits" of matter to explain phenomena. These entities, though unseen, are taking on a reality for them. We might be tempted to view science education as a process whereby pupils' naive conceptions are gradually and progressively shaped toward those of school science. Such a view, however, would also need to account for the important differences that exist between "everyday reasoning" about phenomena and the scientific pursuit. In "everyday reasoning," the criterion for acceptability of a particular model tends to be utilitarian: Does it help get the electrical gadget working, find the fault in the car engine, etc. For scientists, the criteria of coherence and parsimony are more influential.

What pupils and scientists see as appropriate explanations of phenomena differ. Pupils often provide explanations in terms of a linear sequence of events in time rather than a modeling process. These are some of the reasons why viewing school science learning as a continuous process of conceptual evolution may be simplistic. We need to recognize and address the potentially important discontinuities between the kind of reasoning used in everyday situations (to which naive conceptions are adapted) and the formal pursuit of science. Here, investigations into children's views about science itself would be most informative.

A final consideration necessary to gain an understanding of students' general conceptions about science concerns the influence of personal or social experiences on their construction of models. Here, the research suggests two alternative perspectives. The perspective that derives from a Piagetian tradition holds that knowledge of the world comes about through the individual's spontaneous interactions with the physical environment. The perspective is clearly presented by Strauss (1981) who, in accounting for commonalities in naive conceptions, argues:

The common-sense representation of qualitative empirical regularities is tied to complex interactions between the sensory system, the environment that supplies the information...and the mental structures through which we organize the sensory information which guides our behaviors. I argue that individuals' common-sense knowledge about qualitative physical concepts is no different today than in the times of, say, Aristotle. (p 297)

An alternative perspective places greater emphasis on the social transmission and construction of knowledge (Solomon, 1987). From this social perspective, researchers argue that the

mental models used to organize experience are culturally transmitted. (The conceptual environment in which humans live in the twentieth century differs considerably from that of Aristotle!) If science itself, as public knowledge, is socially constructed, then learning science must be seen in terms of a process of social transmission.

Edwards and Mercer (1987) argue this point:

However active a part pupils are allowed to play in their learning, we cannot assume that they can simply reinvent that culture through their own activity and experience. It is necessarily a social and communicative process and one which has as an inherent part of it an asymmetry of roles between teacher and learner. (p 157)

EDUCATIONAL IMPLICATIONS

If we accept that science learning involves the restructuring of students' conceptions, then understanding the processes by which conceptual change occurs becomes a central issue for teachers.

Drawing on accounts of the history of science, Posner, et al. (1982) suggest that a number of conditions must be met for conceptual change to take place. First comes dissatisfaction with existing conceptions, then a new conception that appears intelligible, plausible, and fruitful in offering new interpretations. Other research further suggests that conceptual change can be potentially threatening to the individual (Claxton, 1984) and that students may require a supportive environment in which their ideas are valued if they are to explore new ways of thinking.

Ways of promoting conceptual change in classrooms have been investigated by a number of research groups worldwide (Champagne, et al., 1982; Driver and Oldham, 1986; Hewson and Hewson, 1984; Nussbaum and Novic, 1984; Osborne and Freyberg, 1985). These groups have suggested teaching strategies to facilitate conceptual change. These strategies include:

- providing opportunities for pupils to make their own conceptions about the topic explicit so that they are available for inspection;
- presenting examples that challenge children's prior ideas. Counter examples by themselves do not provide children with new conceptions. They may, however, provoke children into considering the need to rethink their ideas. Children use various strategies to avoid conflicts. They may select and fit observations to their existing ideas or argue that the counter example is a special case;

- using strategies that enable pupils to consider and evaluate alternative conceptions of presented phenomena;
- providing opportunities to use new conceptions. Long-term accommodation of a person's conceptions is unlikely to happen if the new schemes are not seen as useful;
- giving pupils opportunities to become more aware of their own conceptions and how they change. Researchers have studied the effectiveness of various techniques for developing pupils' metacognitive strategies, including concept mapping and personal learning logs.

Research into conceptual changes taking place during instruction in specific topic areas conducted by the Children's Learning in Science Project at Leeds suggests that some commonality exists in the conceptual trajectories that students follow. These can, therefore, be anticipated and taken into account in developing reflexive teaching sequences (Brook, 1987; Driver, 1988), i.e., teaching sequences that account for the child's ideas and progress learning. In order to do this, however, teachers will need to develop strategies for making assessments about the level of children's thinking.

A conceptual-change view of learning also has implications for longer term curriculum planning in science. Developmental studies in specific domains show how children's schemes are restructured progressively over periods of years. This has implications for the long-term organization of learning experiences in school curricula.

IMPLICATIONS FOR ASSESSMENT

Developmental studies of children's scientific performance have implications for assessment. In commenting on this issue, we must consider the purposes for which assessments are undertaken. For many years, educational assessment has been primarily normative in nature. Cohorts of students have been assessed and their performances compared with others in the population. Such assessments have provided information on how an individual's performance compares with population or group norms. We now recognize that this process provides very limited information to individuals about their progress or to teachers about how to promote children's learning in their classes. An alternative purpose for assessment is to provide information about students' science performances and to give an indication of the progress they are making as individuals.

Comparisons are not with other students but with the same student over time. Assessments of this kind, designed to document the progress of individuals, can be informed by developmental studies in a number of ways.

Any scheme to document children's progress in science requires answers to the following questions:

- Along what dimensions do pupils' performances progress? (i.e., What progresses?)
- In what way and to what extent can children be expected to make progress along such dimensions?

Developmental studies relevant to science undertaken over the last two decades have tended to document development in generic thinking skills, mainly informed by Piagetian stage theory. Large-scale surveys have established the proportion of school children at different ages whose performance on tests indicates they are operating at specified stages within the Piagetian models (Lawson, 1985; Shayer and Adey, 1981). Within this perspective, progress occurs when students move up through the stages of thinking. It is also argued that information about the stage of pupils' reasoning skills is useful in making pedagogical decisions since it enables teachers to match the cognitive demand of lessons to the child's stage of thinking.

However, some researchers now question whether progression through general stages gives an appropriate picture of children's progress in science. Studies indicate the extent to which children's reasoning in science is dependent on particular contexts and knowledge domains (e.g., Donaldson, 1978). For this reason, studies of progression in children's understanding within particular domains can provide important background documentation against which to assess progress. However, such studies have been undertaken in only a few areas, and further work of this kind is necessary if such studies are to provide the background information against which the progress of individual students can be judged.

The development of general reasoning strategies independent of specific current domains remains an open question. To investigate this, studies could usefully explore the extent to which progress in domain-specific reasoning in science supports or promotes generalized reasoning strategies and vice versa.

To be informative, assessment schemes should be grounded in research on development in children's scientific reasoning. If we know what aspects of children's scientific reasoning progress, then we can collect information that indicates the progress individuals are making. Such information can be useful not only in monitoring individual performance for the purpose of reporting (e.g., to

parents) but also, for informing further teaching: It indicates learners' present understandings and competencies and, therefore, can guide decisions about the next steps teachers should take.

In practice, assessment approaches that provide valid information about progress will need to account for a number of factors. If we are to assess progress, we cannot judge a student solely on performance outcomes: In some areas of reasoning, performance in terms of correct responses does not progress but competence in the types of reasoning used does. Thus, we need assessment approaches that can probe reasons for responses as well as the answers themselves. Furthermore, if children are to demonstrate what they are capable of in more complex situations, they need time to get into a problem. They need to sort out what may be relevant or not, to process information, and to judge or assess the solution obtained.

In summary, developmental studies of children's science performance could valuably inform and underpin assessment programs. They could provide a theoretical basis for decisions about which dimensions to select for assessing children's progress. In addition, they could provide background information for teachers and educators about the "road map" of progress that can be expected. Finally, if appropriately organized, such assessments could provide teachers with information that can inform further teaching interventions.

PART FIVE

New Approaches to Science Assessment



Introduction

George E. Hein

Between the two extremes of total reliance on testing and using actual experiences for assessments lies a great middle ground of efforts that combine some aspects of the ease and reliability of testing with the authenticity of performance. In describing this broad area, most authors grope for actual examples and deplore the lack of models.

Despite shortcomings of most tests and the difficulties of applying them to individual situations, we have little trouble imagining the benefits that can accrue from testing, or how, in principle, the system would work. Some activity, separate from everyday instruction, is administered to students and the results are evaluated and interpreted to assess knowledge. Similarly, if assessment is totally integrated with practice, then the students would be asked simply to “do” science, to do the everyday work of the classroom, and some judge or panel of judges would assess this work based on agreed upon criteria.

Our difficulties with each of these models make us seek activities that can be called assessment in the sense that they are not only specifically carried out (at least in part) for what they can tell us for purposes of judgment, but also have some clear relationship to “doing” hands-on science. Eleanor Duckworth’s observations of the African Primary Science project were assessments of this kind; Piaget’s clinical interviews have suggested an approach to many evaluators. The call for portfolios or authentic tests, those using an actual piece of curriculum as the assessment, proposed by the Coalition for Essential Schools, are efforts to lead the education community in this direction. But actual, documented efforts along these lines are few.

The two papers that follow illustrate two different approaches, each of which makes use of children’s “natural” school activities specifically for assessment. Edward Chittenden has studied young children’s growth and development for years, primarily the development of early literacy. He describes what we can find out about children’s knowledge of science from ‘staged’ classroom conversations. The method is not simple and requires more resources than would be needed to give a test, but the paper illustrates the wealth of information for instruction, curriculum

design, and school policy that is generated from this semiformal assessment method. Hubert Dyasi's paper provides a parallel discussion based on the analysis of children's drawings and writing.

Each case requires us to develop an analysis scheme, collect assessment data systematically (although the actual material is the product of 'normal' classroom activity!), and apply a consistent strategy to its interpretation. The papers also demonstrate the richness of this approach, the potential power of collections of naturalistic data, and, above all, the applicability of these methods in the classroom. Individual teachers can use these methods profitably both to track children's progress and to assess how a class is progressing through the year.

In addition, the work described in both papers illustrates that these alternative assessment methods can be of value to schools and school districts. The ideas that children bring to discussions, the pictures they draw, and the stories they write inform administrators of their concerns and their backgrounds, of the knowledge that is available in that community or school, and even of the particular mistaken notions that may prevail. This information is essential if we wish to provide resources to classrooms to develop locally appropriate science curricula or to adapt national curricula to local conditions.

Young Children's Discussions of Science Topics

Edward Chittenden

BACKGROUND: DISCUSSIONS AS A SETTING FOR ASSESSMENT

Primary science is often characterized as hands-on to the extent that young children are engaged in active investigation of physical and natural phenomena. The raw materials of science—animals, plants, balances and magnifiers, sand and water—are resources for the program. Equally, however, primary science can be characterized by the “hands-off” dimension of social interaction among the children, in the form of their conversations, discussions, and talk. Ideas, along with materials, are brought into the classroom, to be shared, examined, and reconsidered in various ways. The social quality of children's learning in science is well established in the theoretical literature, from Piaget (1950) to Vygotsky (1962), just as it is appreciated in practice by observant teachers (Schwartz, 1986; Strieb, 1985).

Although most discussions are informal and spontaneous, teachers also promote conversation through the more formal settings of group discussions and “class meetings.” These are occasions when the children come together specifically to talk about recent experiences and observations related to science activities. Teachers may use such times to introduce something new as well for review and reflection.

Aside from their instructional value, class meetings can provide occasions for assessment. The give-and-take of open discussion can provide the teacher with cues about children's thinking. Certain ideas or questions may persistently come to the surface, enabling the teacher to evaluate children's understanding better and to plan accordingly.

The study described here attempted to capitalize upon group meetings as a setting for investigating children's thinking. One purpose of the study was to identify patterns in children's interests and ideas that could serve as a framework for evaluating science books and instructional materials for this age group. A second, methodological purpose was to explore the value of group discussion settings as contexts for assessment of children's thinking. We selected three areas of primary science: shadows-light-reflection; earth-sun-moon; insects.

Edward Chittenden is a research psychologist at Educational Testing Service, Princeton, N.J. In collaboration with elementary school teachers, he has investigated naturalistic approaches to assessing children's learning in school.

PROCEDURES: DATA COLLECTION

Twenty-two teachers, kindergarten and primary levels, from nine different schools participated in the data collection. These teachers, along with 20 additional practitioners (teachers and curriculum specialists), also participated in data analysis. The schools, eight public and one private, represented urban, suburban, and rural areas.

Guidelines for Discussions

In collaboration with the teachers, we developed a set of guidelines for conducting discussions. In format, the discussions generally followed a class-meeting format. The guidelines were designed to introduce some procedural uniformity across the classrooms, yet allow teachers the flexibility that is essential to a naturalistic method. The guidelines included the following:

- that discussions begin with open-ended questions, such as:
 - What have you noticed lately about our caterpillars?
 - What are some things you know about shadows? What is a shadow?
 - What sorts of questions do you have about the sun? What have you wondered about?
- that teachers refrain from correcting or unduly modifying the children's comments
- that discussions proceed in a manner ensuring the involvement of most all of the children
- that records be made of each child's statements

The guidelines were intended to encourage the sort of discussion that is sustained by child-initiated questions and ideas, and that allows the children some control over the direction, or drift, of their conversation. The teachers' open-ended questions and low-key role were intended to bring out evidence of the children's own interests, expressed in terms of their choosing. The requirement that most children should be assured a place in the discussion was intended to offset possible dominance by a few.

The fourth requirement, recording, was handled in a number of ways. In some classrooms, an observer took detailed notes of the proceedings; in other classrooms, discussions were taperecorded and later transcribed. In still other situations, the teacher made notes during the discussion. Given practical restrictions on data collection, most of the records that were eventually

collected did not represent complete transcripts of the children's language, but they did contain the children's key words and phrases.

Topics for Discussion

During the planning phase of the project, several meetings were held with participating teachers to identify science topics that appeared most promising. Criteria for topic selection included: high interest to children; pertinence to the primary curriculum; and diversity within the domains of scientific inquiry represented. Selection was also guided by a review of existing books and instructional materials for children. For some topics, the science literature for young children is fairly rich (insects, especially butterflies), while for others (shadows/light), it is sparse.

Eventually, we identified three topic-areas: shadows/light/reflection, earth/sun/moon, and insects (with an emphasis on caterpillar/butterfly life cycle). Other topics we considered were rocks/minerals/dirt, electricity/magnets, seeds/plants, and molds/rotting. In the interest of making the project manageable, we restricted the list to three. (We did append a miscellaneous category to accommodate some additional interests.)

Decisions concerning the timing of the discussions and choice of a particular topic-question were up to each teacher. Generally, teachers selected topics and questions they felt were most relevant to their ongoing program. Insect discussions, for example, could be timed to the appearance of caterpillars in the classroom; astronomy discussions might take place during a month when the moon and sun were topics of special interest.

Participating teachers recorded a total of 75 discussions, with durations ranging from 10 to 40 minutes. The modal number of statements in a given discussion was around 30, somewhat less for kindergarten than for first or second grades. All statements were eventually entered into a microcomputer data base program (Reflex) to be organized and printed, in raw form, for purposes of data review.

Participating Teachers

The classrooms of the participating teachers were characterized by an activity-based approach to science instruction. In these classrooms, the raw materials for children's "hands-on" investigation were much in evidence. The classrooms were also characterized by a fair amount of talk and conversation, among small groups of children or within a larger group. Given these classroom qualities, the discussion guidelines were congruent with

classroom routines and practices. The major difference, at least from the perspective of many of the teachers, was that the project's discussions were more "staged" and the teacher's stance tilted toward "listening and observing."

PROCEDURES: DATA ANALYSIS

Descriptive (Emergent) Coding: Themes and Patterns

We initially examined discussion records for evidence of themes and patterns in the children's comments. Through a series of working conferences, teachers, curriculum specialists, and research staff undertook a close study of a representative sample of the discussion protocols. In these meetings, participants worked in different teams (six or seven people per team) for each topic area: astronomy, insects, and light/shadow.

Drawing upon procedures developed at the Prospect Center (Carini, 1975), each team described the selected discussion in considerable detail. The process entailed reading and re-reading a full discussion record (consisting of perhaps 20 to 30 statements) for evidence of patterns in the children's remarks. Typically, children picked up and elaborated certain ideas, while ignoring or discarding others. Similarly, certain images or phrases might have been re-stated by several children in different ways, suggesting their salience above other formulations.

From close description, initial themes or "headings" were identified that highlighted different aspects of the children's statements; these headings were used in subsequent analysis of the remaining samples. Of particular interest were issues or ideas that seemed robust, emerging in discussions across several samples. For example, discussions about the moon often turned, at some point, to questions about "what happens to the moon in the day-time?" And discussions of the earth in space had a way of provoking large questions about origins, ancestors, and the inside of the earth. Similarly, shadow/light discussions brought out observations or ideas about a shadow's connection to the person. These sorts of issues and images emerged whether the discussion took place in classrooms in central city Philadelphia or rural Vermont.

Categorical Coding: Forms of Statements

The second approach to data analysis considered the form rather than the content of children's statements. That is, we distinguished between the different ways that children framed their statements: question, assertion, observation etc.

A coding scheme built around five categories of statements was developed to be applied across all discussions. The first type

of statement, termed *pronouncements*, took the form of relatively straightforward, unembellished declarations of fact or “bits” of knowledge. Pronouncements provided few cues about the speaker’s sources of information. They also tended to carry an assertive, nonspeculative tone:

- The sun is hot boiling gases.
- Fire gives you light.
- Centipedes have 23 legs.

A second form of statement, *observations*, consisted of reports of personal observation. In these statements (which sometimes carried a quality of pronouncement), children revealed something about their sources of knowledge as they referred directly, or indirectly, to specific observational experiences. Many statements were recollections of experiences outside the school; others referred to the classroom environment:

- When you get up early you can see the moon.
- I saw a rainbow at a waterfall.
- It (butterfly in classroom) has eight feet. I counted. Four on each side.

The third major category, *speculation/explanation*, represented children’s explicit efforts to formulate complex explanations that took account of multiple variables, attempted to deal with cause and effect or, perhaps, juggled their classmate’s comments:

- Three things make shadows: yourself, light and dark.
- It’s (earth) real hard. The earth was rounded by air shoving against it.
- The stomach part of the butterfly is what is left of the caterpillar.

The fourth category, *question*, contained the full range, from children’s requests for specific information—“Is the flag still on the moon?”—to inquiries into larger problems:

- Why do bugs stick?
- When the earth was made did the first elephant gradually come out of the ground?

A fifth and smallest category, *references*, identified any comments in which the child specifically cited secondary sources of information. Examples include references to books, TV, museums, or comments such as, “my brother told me.”

Although the majority of statements could be classified within one grouping, the coding categories were not mutually exclusive, and the boundaries between categories are admittedly

fuzzy. For example, a child might well embed a specific observation within a formulation or question. Also, the discussion context needed to be taken into account when coding a particular statement. Thus, statements that would appear to be pronouncements if viewed in isolation might well have been considered explanations when the comments of the other children were considered.

RESULTS I: GENERAL QUALITIES OF CHILDREN'S DISCUSSIONS

In general, the children's discussions displayed qualities of cohesiveness and focus. The children's comments, for the most part, were addressed to the subject at hand. And although the topic or agenda might shift during the discussion, such changes grew out of a line of thinking. Even some of the comments that seemed "off the wall" when uttered turned out, in retrospect, to be the child's way of entering the problem.

The children also indicated that they were listening to one another when they expanded upon a classmate's comment or restated the issue in their own terms. In a number of the discussions, the topic became differentiated and elaborated; what started out as a series of "facts" and pronouncements moved toward speculation, hypothesis, reexamination. Arguments and attempts to correct a classmate's comments were relatively rare. The children tended to let conflicting statements coexist, even when the contradictions seemed blatant (at least to the adult).

In noting the cohesiveness in discussions, we do not claim to have undertaken systematic analysis of social discourse, as this was not the intent of the project. However, in light of stereotypes about young children's limited capacities for attention, the sustained nature of their participation needs recognition. Undoubtedly, this speaks to the skills of the teachers and the traditions of their classrooms as much as to the abilities of the children. Moreover, in the opinions of teachers, the nature of the topics themselves may well have fostered conversation. "Science" questions invite broad participation and interaction. As the following examples illustrate, shared experiences, ideas, and teachers' focusing questions supported the cohesiveness of children's discussions.

Shared Experience

In some discussions, shared experiences provided the basis for the cohesiveness or connectedness of children's comments. For example, discussions of shadow and light often brought out accounts of personal observations. These were the sorts of observations:

- When I swing, when I go up the shadow gets bigger.
- When I run, the shadow in back of me goes faster and gets in front of me.

Similarly, a class's discussion of insects brought out observations that were shared and connected, as in the following statements about the Monarch's chrysalis:

- There's a crack at the bottom.
- On the side there's a little tiny part where it's gonna come open.
- It looks like you can see the wings.
- I can see some of the white dots.

Common Themes

In other instances, the discussion was focused in that certain themes or ideas came to the fore and remained as strands in the conversations. In response to the teacher's question, "What are some things you know about the earth?" a first-grade class discussion started out with the following sequence of statements offered by the first five children:

- The inside of earth is hot.
- There's no end to the earth.
- Astronauts discovered the earth.
- Some of the earth is made of water instead of cement.
- Long long ago earth was made of water.

This discussion continued for 25 minutes. What is of interest is that the opening comments quoted above carried several themes that were subsequently pursued by others, e.g., the earth's composition; the earth's origins; and the earth as an object in space.

Group Problem Solving

Sometimes the discussions became centered around complex problems, leading to attempts to develop explanations and theories. For example, in one urban first-grade classroom the children talked about the heat of the sun and the ways they had experienced it. (This discussion took place during a period of particularly warm weather in June.)

- If you touch a wall, you can feel it's hot.
- If you have shorts on, you can't sit down in the car because the seat's too hot.
- Even at night it's hot.

One child then raised a question that caused a pause in the conversation:

- If the sun makes it hot how could it be hot at night, even though the sun has gone down?

After some further comment about how hot it could be at night, another child formulated a hypothesis that was generally accepted as plausible by others:

- The sun reflects off the moon so there's going to be some heat left in the moon...if there is a lot of heat in the moon then it will be warm at night...if a little heat then it will be cold.

Many of the discussions observed in the study allowed the children to open up a topic and consider its facets. Each speaker contributed certain thoughts or ideas, enabling the children, collectively, to consider the matter in its greater complexity. Thus, a discussion that started out as a series of pronouncements or reported observations could elicit more complex statements as the children centered upon some issue or problem.

RESULTS II: TOPICS AND THEMES

The following sections describe the results from coding for each of the three topic areas. Patterns and themes in the children's discussions, discernible across classrooms, are illustrated.

Shadow/Light/Reflection

Phenomena of shadows, light, and reflection are familiar to children. When children talked about these matters, they were not so much discussing "science" as talking about everyday happenings and experiences.

The shadow/light discussions produced the highest proportion of reports of personal observation of any topic. Approximately one-third of all statements contained fairly explicit reference to something seen and experienced (table 1). Children recounted their observations of such matters as the movement and shape of shadows; the play of shadow and light upon bedroom walls and under street lights; and reflections on shiny surfaces of parked cars and puddles. Such observations and comments tended to be accepted as true and familiar by the group; the discussions seemed to tap a common core of childhood experience. In the same vein, children rarely asserted an "expert" opinion, as they sometimes did in astronomy discussions, presuming greater knowledge than their classmates.

A common pattern in these discussions took the form of a succession of statements, each child adding some idea to the central topic. The statements were often prefaced with the conditionals "when I..." or "if you..." which established the context of

TABLE 1. Characteristics of Children's
Statements in Discussions of Three Science Topics

	TOPICS		
	Earth Sun/Moon	Insects	Shadows/ Light/ Reflections
Total # Discussions	27	23	15
[Total # Statements]	[1008]	[493]	[419]
[Average # Statements]	[37.3]	[21.4]	[27.9]
Categories	%**	%	%
Pronouncements [facts/assertions]	54	45	43
Reported Observations (Classroom-based)	16 (1)	23 (18)	35 (8)
(Outside classroom)	(15)	(5)	(27)
References	7	4	2
Speculation/Explanation	28	24	17
Questions	16	20	6

*Categories are not mutually exclusive

**% Total # Statements

the observation and imparted a story-like quality to the account. As illustrated in the examples below, settings were often specified:

- When I play rope, I can see the shadow of the rope moving.
- Shadows are made out of light—I've seen shadows when we go to the mall, under the restaurant light.
- Trees have a shadow...I've seen shadows of leaves on the window.
- When my brother gets back into bed, I see his shadow.

Some of the statements suggest informal experimentation on the child's part:

- When I run, the shadow in back of me goes faster and gets in front of me.
- I flashed the flashlight to the mirror and the flashlight shone back on the wall.
- You can see yourself in water, but it comes upside-down.
- The closer you are to the wall the bigger the shadow.
- The television makes light if you put your hand out.

Shadow and Light

The children used the three key terms *shadow*, *light*, and *reflection* freely, but attention to shadow predominated and generally constituted the point of departure. In the children's statements, shadows seem more tangible, less elusive, more under control of the person. When teachers began by asking, "What do you know about shadows?" the ensuing conversation tended to be longer, and perhaps more focused, than were discussions prompted by an opening question concerning light. As one child remarked, "Light's just light."

What Makes a Shadow

The following sequence of statements, taken intact from one discussion, begins and concludes with two very different statements about light's connection to shadow. The discussion moves from a listing of the ingredients needed to make shadows to the idea of blocking and the absence of light. The excerpt, which retains the original sequence, is an example of the qualities of elaboration in class discussion noted earlier in Results I.

Question: What makes a shadow?"

- Light
- Dark
- Light and dark make shadows.
- Yourself
- Three things make shadows: yourself, light, and dark.
- You, lightness and darkness.
- You block the lightness and it makes darkness.
- A little bit of light and a lot of dark
- Mostly yourself
- Everything in the world makes shadows except light...so light doesn't make shadows.

In addition to making the general assertion that light was needed to make shadows, some of the children talked about the position and shape of shadows in relationship to the sun. Comments of this sort were noticeable in classrooms where teachers had introduced related playground activities. One child summed it up by saying, "We are like sundials." Others added, "Light hits you and makes a shadow," and "We traced shadows (on the playground) and later the sun was out of line."

As illustrated in the above segments, children introduced the subject of light in connection with shadow. Light is the background, shadow the figure. They view light as a necessary, essential condition:

- You need light to see.

- It (light) is something that makes people see. You don't have to feel for things.

When children were asked to talk about light as such, they tended to speak in terms of sources:

- Light is fire.
- Light is like lamps.
- It comes from the sun.
- I think light is electricity, that's how light bulbs work.

Reflection

The topic of reflection was sometimes introduced by the teacher and sometimes raised by the children within the context of discussions of shadow. As they could with shadows, children readily recalled situations in which they had observed a reflection. For the most part, reflections were depicted as a variant of shadow, differing in important respects but also similar in essential ways, such as connection to person. Children reported that: shadows are black or dark, but reflections are colorful; shadows follow you but reflections are less predictable; you can find shadows almost everywhere, but you need something shiny to see reflections:

- A shadow is your reflection without the color.
- A shadow shows you a picture of yourself and you don't have to look in the mirror.
- Shadows are a little bit like reflections...reflections are usually colorful and shadow is plain dark.
- A reflection comes from mirrors, metal, glass...a shadow comes from stone and wood. Things block out the light and there is only dark.

Conjecture, Theory, and Question

Compared to conversations about astronomy and insects, discussions of shadow and light were less likely to move toward speculation and question (table 1). The "everyday" phenomena of shadow and light seemed taken for granted. There were relatively fewer attempts at setting forth a theory or elaborated explanation, and fewer spontaneous questions. The children seemed comfortable with their knowledge, even when they voiced what appeared to be quite contradictory opinions or observations.

Nevertheless, discussion did sometimes lead to the kind of question that teachers could have used to promote a more critical, investigative stance. In the midst of one discussion a child suddenly asked, "What happens if you hold a mirror to your shadow?" and another speculated, "If you were standing on your head you'd have a little shadow."

Also embedded in many remarks was an element of mystery and illusion—a child's sense that things aren't always what they may seem:

- Sometimes you seem bigger than your shadow...sometimes smaller.
- Little things can make big shadows.
- Clouds are like shadows because shadows follow and clouds seem to follow.
- Shadow can't be two miles behind looking for the person.

Sources of Information and Ideas

References to secondary sources of information, such as books about shadows or TV programs, were rare in these discussions. In three different classrooms, children did recall trips to a science museum, remembering especially the "shadow" room in which person and shadow are separated:

- We got to walk around in that room and make shadows. You stand in front of a light, making a shadow on the wall, and when you walk away, your shadow stays on the wall!

Earth/Sun/Moon

The tone of the earth/sun/moon discussions was very different from the everydayness of the shadow discussions. As shown in table 1, these topics were characterized by the highest proportion of pronouncements and speculations, and the fewest instances of reported observations. The astronomy discussions also contain more frequent references to secondary sources of information.

Despite limitations on their understanding, young children are clearly interested in such "large" questions as the earth's movement, the patterns of day and night, the appearance of the moon, and the power and light of the sun. The fact that astronomy discussions tended to be the lengthiest of the three topic areas is but one indication of their persistence (table 1). Another indication is that children stayed with questions of this sort; astronomy discussions did not, in general, turn into space travel fantasies.

A more substantive sign of interest is found in pronouncements and questions indicating that children had given these matters some prior thought. As shown in table 1, astronomy discussions were characterized by the highest proportion of speculative comment and assertion. The children raised large, nontrivial issues, often posed with a tone of seriousness. Some of these issues, such as the earth's movement, were pursued in the conversation; others, such as proclamations about the biggest planet or

the size of stars, tended to be left standing, as if stated for the record.

Some examples of "large" statements, excerpted from several classrooms:

- A lot of stars that are really huge blow up.
- Why is the sun so bright because space is so dark?
- When the world turns, why doesn't everything fall down?
- Ancestors are old planets.
- If you stay in the sun for 5 hours you might get killed.
- Why is the moon in two shapes when the sun is in one shape?
- The earth is a gigantic asteroid.
- There's something deep inside the ground, it's a hot ball of fire.

The discussions of earth/sun/moon also suggest that a children's science culture exists around these topics. Some of the matters raised in the discussions were very likely the subject of informal conversations among children, both within and outside of the classroom. And, it is no accident that siblings were cited as authorities as frequently as were parents or teachers, raising the possibility of ideas moving through generations of childhood. Interestingly, some statements made by these children of the 80s were "recognized" by their teachers as ideas from their own childhoods:

- If you dig through you get to China or Italy.
- If you look at the sun your eyeballs will burn out.
- If the earth turns over at night, why don't I fall out of bed?...sometimes I think this.

Themes

Discussions were analyzed for themes associated with earth, moon, and sun. Not surprisingly, each of the three subtopics elicited distinctively different images and questions. The moon, for example, tended to be talked about in a more personal manner; children pursued questions of what it did in the daytime (behind trees, clouds, on the other side of the earth) or whether it followed you. They also voiced a recurring interest in "where it got its light."

Discussions of the sun were marked by reference to its powers and effects. Children talked about its awesome composition ("hot boiling gases," "acid," "one huge atom"); they described it as very bright, shiny, and as extremely hot and potentially dangerous ("You can get skin cancer." "It can make you blind."). They also spoke of the sun as the source of light ("for all the world") and as a necessity of life ("It makes flowers grow.").

Images of Earth

The moon and sun are comparatively easy to talk about in the sense that they are objects at a distance to be looked at and thought about. Although their apparent movements and changes in appearance raise issues, they are nevertheless apart from the observer. By contrast, earth as object is elusive. It is something we live upon, yet (unless we are astronauts) cannot observe except in close-up specific fashion. Moreover, in common parlance, the term *earth* is open to various meanings and uses.

At least five or six quite different meanings of earth can be found running through the children's remarks. These meanings are outlined in table 2 and summarized below. They are not meant to represent different conceptual models of earth, but rather different views, determined in part by the context of the discussion. (Most likely, the children juggle several pictures of earth, rather than adhere to one.)

- Earth as an object in space. The focus of these remarks is upon the earth's position or movement in space. The observer seems detached, as someone watching this object from afar. Some of the younger children claimed that you can look up and see the earth in the sky. One stated, "that's why we have to go up in space." When asked by the teacher to point toward the earth, some children pointed up.
- Earth as the ground, surface, dirt. In these remarks, the earth is depicted as the ground "we walk on," although there was some debate over whether a sidewalk is really the "ground" or whether ground consists of dirt. There is corresponding interest in what the earth is made of: cement, water, mud. When asked to point to the earth, some children pointed down.
- Earth as interior, beneath the surface. The heart or essence of the earth is considered to be "inside," beneath the ground. Sometimes the term "core" is used and on more than one occasion a child seemed to infer that if the earth has a core, it also has/had a seed ("Did the earth grow?"). The children invoked images of lava and volcanic forces and talked about the power and force of gravity.

TABLE 2. Where Is Earth?

The earth in space (up in the sky/part of the solar system)

- "up in the sky"
- "no, not up in the sky, in outer space"
- "because the earth is in outer space"
- "The earth is in space--that's why we have to go up in space"
- "Earth is a planet."
- "The earth is a gigantic asteroid. Lots ofthat burns"
- "The earth when you look up in the sky you see blue. . you see the earth in the sky and the earth is around you.
At night you see the stars and it tells you the months. Like December's up in the sky."

The earth is all around us

- (points sideways, then up--Q: Why pointing both ways?) "Because I was thinking there was the earth around us."
- "We live in the world. On the earth."
- "It's all closed up. There's no place to get out, they've closed everything up."
- "I think the world is on the earth."

The earth as the ground or surface

- "We're on the earth."
- "We live on the earth--got grass."
- "When you stand on the earth it feels hard."
- "We're on the flat part of the circle."
- "Some of the earth is made of water instead of cement."
- "The continents are pulling apart and making the ocean deeper and longer."

Interior/under the earth (essence is inside)

- "There's something deep inside the ground. It's a hot ball of fire."
- (points down) "Because the earth's core is down" (Q) "It's the heart of the earth."
- "The earth is under the ground. When you walk on the ground, and the earth is underneath."
- "It could explode like volcanoes."
- "Around the earth is very hard and underneath is not so hard."

Earth is represented in models (globes/maps/pictures)

- "One of the States is shaped like a boot."
- "I've seen a picture of the earth."
- "I know the earth is round because I have an earth in my house."
- "I don't believe she has the earth in her house, the earth is big. She's living on the earth. It's a globe."
- "Earth is round, I see it on t.v."
- "How did states get on top of the globe?"

The earth as a place to live (habitat)

- "lots of life forms on earth"
- "There are a lot of different places on the earth."
- "The earth got people."
- "The earth got animals."
- "There's pretty flowers."
- "There's grass and food so you don't die."

- Earth as all-around-us—the world. From this perspective, we live in the earth, not on it. In response to the pointing question, some children extended their arms horizontally, “because the earth is all around us.” The boundaries of the earth seem to surround its inhabitants.
- Earth as the globe, map, picture, photo, TV image, diagram. In these statements, children referred to different conventions for representing the earth. They sometimes cited these representations as the principal source of evidence to support opinion concerning the earth’s color and configuration. In some statements, children seemed to equate the model with the thing represented: “I know the earth is round because I have an earth in my house.”
- Earth as habitat, as a nice place to live. These statements tended to portray the earth as a dwelling place where people and animals live, surrounded by trees, houses, etc. Earth is generally depicted as a benign and pleasant environment.

Compelling Questions About Earth

Children have apparently encountered, and attempted to assimilate, certain “facts” about the earth; they regard these facts as important and in need of discussion. Examples of two matters that regularly sustained discussion are the shape and the movement of the earth.

The subject of the earth’s shape was raised in almost every discussion. All children agreed that it was “round,” but they displayed different ways of reconciling this fact with their experience:

- We’re on the flat part of the circle.
- The earth is flat...but it’s round.
- We don’t know that we’re tilting. The earth is round, so we’re in a shape like this...my farm is this shape. Only I don’t know it.

Movement was a particularly strong earth theme, running through 25 percent of all statements pertaining to earth. The ambiguity of movement was suggested in the children’s choice of verbs and attempts at explanation: The earth was said to “go around,” “move around,” “turn around,” “turn over,” “spin,” “roll over,” “twirl,” “rotate.” And children asked questions such as: If the earth is moving, why don’t we feel it? Is it going fast or slow? Where is it headed? And, how does the earth go around the sun, when the sun sets?

One of the more articulate second-grade children explained:

- At nighttime, when the earth goes around, it takes like a year. A day is not a year but it goes around and it seems like it's not as long as it should be.

Taken on the surface, statements such as these convey the child's confusions about the day/night cycle and the orbital year. What is not conveyed in the statement, however, is the energy and effort that goes into such attempts at explanations. Hand gestures and props were sometimes employed as children talked about the earth's movement. Teachers and observers felt that in many instances, the children's language failed them. Indeed, children sometimes employed one term when they apparently meant another. This is not to claim that they "really" understood rotation, revolution, etc. but to suggest that the statements are best interpreted as an indication of the direction of their thinking, not as an adequate representation of their concepts.

Other major themes pertaining to earth were origins, composition, and place in space.

Sources: Observation and References

As indicated in table 1, astronomy discussions contained the fewest instances of reported observations (16 percent). These observations tended to be restricted to three sorts of phenomena: (1) the appearances of the moon; (2) the heat and light of the sun; (3) the movement of the earth. Although reports of observing the moon and sun are not surprising, neither we nor the teachers anticipated the observations concerning the earth's movement. Yet on several occasions, in different classrooms, children became involved in debates about whether or not one could actually feel the earth move. A number of children approached this issue empirically, citing personal experience as grounds for their opinions:

- When you spin around you feel the earth is spinning with you.
- I know the earth turns, because I see the clouds moving. But the earth moves, not the clouds.
- In my basement when I swing and stand up I'm dizzy and it looks like the house is turning around.
- It moves so slow you can't even feel it.

More so than other topics, astronomy discussions included explicit citations of authorities and secondary sources of information. Books and other children, especially siblings, cousins, and friends, were the two most frequently cited; TV and science museum displays were also mentioned a number of times:

- My sister said that the planet takes a bite out of the moon.
- You know what? My cousin thinks there's no such thing as a planet up in the sky.
- I know (earth is round) because I have an earth in my house. It has yellow and green.
- The earth is a circle. I looked in a book.
- I saw stars on the ceiling in the museum.

Insects

Of the three topic areas investigated, children's discussions of insects were the most likely to be associated with classroom activities of observing and recording. During the year, various bugs were brought into the classroom, including spiders, "worms," and grasshoppers. In about half the classrooms, the life cycles of Monarchs and/or silkworm moths were studied. Insects constituted a presence over weeks and months that attracted attention and became the subject of conversation, writing, and drawing. Teachers often began discussions by asking children for comments or questions prompted by their observations.

Children's Previous Experiences with Insects

In some classrooms, discussions were recorded at intervals, permitting comparison of initial discussions, when insects were first introduced, with later discussions. The initial discussions, especially, point up variation among children in extent of background knowledge and prior interest in the topic. Thus, in any classroom, a few children were apt to have knowledge of "bugs" that was extensive, almost encyclopedic, if not completely accurate. For them, insects mattered well before the creatures' introduction into the classroom. Sometimes these children spoke as authorities:

- A spider isn't a bug you know; a bug has at least six legs but no more.
- Once a bee stings someone it can't sting again. If a mosquito does, it can sting again.

Yet for most of their classmates, the terms bug or insect—initially, at least—seemed to cover most any small animal that flies, crawls, or stings. In one kindergarten/first grade classroom the teacher asked: "What do you think an insect or bug is? Can you give some examples?" The children's list ultimately included worms and caterpillars (viewed as synonymous), spiders, stinging yellow jackets, snakes, and guinea pigs:

- Ants are bugs.
- Insect is the kind of animal that crawls around the place.

- A bug is a bee; it will sting you; it will get worse if you keep bugging it.
- I think butterflies live near where bugs do.

Compared to evidence in the astronomy discussions, the children as a group came to the topic of insects with fewer preformulated large questions or pronouncements. And, when compared to the shadow/light/reflection discussions, the evidence of shared everyday experiences was not as strong, the one clear exception being the apparently universal childhood experience of “getting stung.” As insects were actually observed and studied, however, such contrast with other topics shifted. Complex theoretical issues were indeed raised by the children, and discussions became grounded in their shared investigations.

Speculation: Grounded Theory Concerning the Chrysalis

Table 1 indicates that the insect discussions were marked by speculation (24 percent) and questioning (20 percent) comparable in frequency to the astronomy discussions. As noted, much of this conjecture was prompted by children’s direct observations of the insects in their classrooms. These discussions therefore provide examples of children’s attempts at theory development, grounded in the data of their observations.

The period in insect development that provoked the most speculation was the *pupa* (or chrysalis) stage. The question of “what’s going on inside” was raised and pursued in several classrooms. Children had observed the successive molts of the caterpillar leading to the emergence of the chrysalis, and they anticipated the eventual reemergence as a butterfly/moth. The mystery of what was happening in the chrysalis understandably provoked comment.

Two broad classes of problems attracted attention. The first concerned the logistics of the insect’s escape from self-imposed confinement. When considering this problem, the children were not so much dealing with metamorphosis as they were with the insect’s predicament. Their “getting out” theories revealed knowledge of caterpillar behavior: They mainly proposed that the insect would eat (chew, nibble, spit) its way out. Some also proposed that the insect would get out by force—by pushing with its legs or newly developed wings:

- First it nibbles a hole and then pushes its way out.
- It’s easing its way out.
- Maybe the wings are trying to bust out of the chrysalis.

A few children noted the constraints of confinement:

- I wonder about how it changes in such a small place, when it's a big butterfly when it comes out.

The second type of problem, drawing greater attention (approximately two-thirds of speculative statements), concerned the nature of the metamorphic process, the transformation from caterpillar to butterfly. Children's prototheories invoked mechanisms and metaphor. The least elaborated hypotheses suggested that the insect accomplished the change through sheer exertion or, alternatively, through a kind of hibernation:

- It kept working and working without eating or drinking, working into a butterfly.
- Maybe he's sleeping and changing.

Other comments suggested that the caterpillar was more or less "conserved;" the process of change being incremental as new appendages were added and/or old ones modified:

- He's starting to grow wings.
- His stomach is getting fat.
- Its antennae are getting longer.

Changes of a more complicated kind are implied in children's statements to the effect that certain parts of the caterpillar become restructured or otherwise transformed into recognizable sections of the butterfly. Some examples:

- The stomach part of the butterfly (abdomen) is what is left of the caterpillar.
- I think that the body of the caterpillar is the middle of the butterfly. 'Cause it couldn't just change.

Perhaps the most complex analysis was offered by a child who drew an analogy to fund raising, an analogy that is scientifically appropriate if one thinks of the caterpillar as investing in protein:

- The caterpillar fund-raises to help the butterflies come to life. Like sends in money and supports it and stuff. It supports it and makes it. If the caterpillar was going to make another caterpillar, he would lay eggs.

Life Cycle and Stages

The concept of "life cycle" is commonly cited by way of curricular justification for the study of insects in the classroom. The four stages of metamorphosis—egg, larva, pupa, adult—provide the framework for much of the instructional material. Most children's books, for example, are organized around this schema, telling the story of development.

A review of the discussions indicates, however, that life cycle and stages, as such, were not compelling abstractions. The questions the children raised and the issues they pursued (i.e., what's happening in the chrysalis) suggest other lines of inquiry and formulation. The "egg stage," for example, is not necessarily perceived as the first step. Instead, life begins with the caterpillar's hatching, and "real life" commences with the emergence of the adult form, as outlined in the following account from a child's journal:

- A caterpillar eats its egg right after it is born. And then it eats the milkweed. Then he gets big and fat and then it hangs like a "J" for a while. Then it come out, like a butterfly. Then it has its life.

The elusiveness of the idea of cycle is also revealed in comments, mainly from younger children, suggesting that the butterfly could return to its cocoon, as to a nest from which it hatched. Children also occasionally referred to a newly emerged butterfly as a "baby butterfly."

Judging from the children's comments, much more can be said about a caterpillar than, for instance, that it represents a phase in the insect's development. Thus children talked about individual differences among the caterpillars; the question of why some were "fatter" or "bigger" than their "litter-mates" was raised in several classrooms. Children also noticed small details in the insects' appearance and habits, and raised questions that may or may not be dealt with in books. In one classroom, the teacher asked each child to pose a question concerning the caterpillars, which had just arrived. Although some children asked the expected, "What kind of butterfly will it be?" most were interested in other features:

- How do they breathe with the top [jar lid] on?
- Why did one caterpillar get fat?
- Why do they all eat the same kind of leaf?
- What are the black things on their faces?
- What is the black stuff in the cage?
- How many pounds do the caterpillars weigh?

Observation

Insect discussions were marked by children's accounts of first-hand observation. As indicated in table 1, almost one-fourth of all statements carried explicit reference to something observed. On some occasions, children recalled encounters with insects outside the classroom. In these statements, they tended to locate the insect within a particular time and space. Locating an observation within a setting is a quality that we also noted in children's reported experiences with shadow and reflection. Here are two

responses to a teacher's question, "How many of you ever found a caterpillar?"

- I found one in my old school on the ground.
- At my aunt's house there was this little caterpillar and he went on my shoe.

The great majority of cited observations referred to insects in the classroom. Teachers sometimes started discussions by asking children to comment on anything they had lately noticed. At other times, the children interjected observational evidence as the basis for an assertion or as stimulus for a question:

- On the antenna there are little dots.
- It has eight feet. I counted, four on each side.
- Why are some of our caterpillars smaller than the others?

In making sense of their observations and those of their classmates, children drew upon their own life experiences for metaphors. They invoked images of eating, drinking, and sleeping. They also made analogies to the experience of dressing, capturing the essence of the action:

- It's splitting its skin, sort of like a zipper (regarding a molting caterpillar).
- It's like stepping out of your pajamas with another set of clothes on underneath (regarding the emergence of the chrysalis).

Sources of Knowledge

Figure 1 depicts several pages taken from a class book that was compiled by the teacher, based upon children's drawings and language as recorded in discussions or journal entries. The children's close attention to detail combined with a flair for metaphor are evident in these pages. In her introduction to this book, the teacher (Kanevsky, 1987) points up the breadth of resources that contributed to the children's thinking over the course of weeks and months. At the center were opportunities for first-hand observation of the successive stages in the Monarch's development. But in addition, films, books, and discussions, integral to development of ideas, informed the work reflected in these pages.

IMPLICATIONS FOR ASSESSMENT

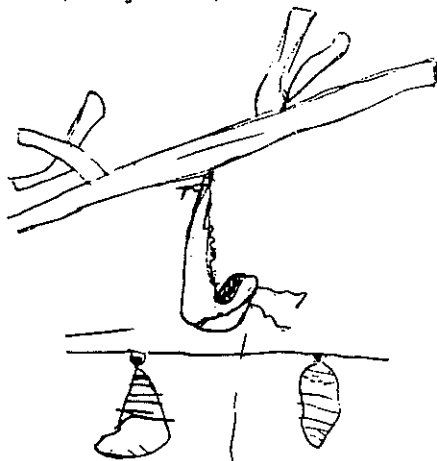
Staging Discussions

Discussions of the sort recorded in this study constitute a sampling of children's conversations about science topics. As an assessment strategy, the discussions can be thought of as "staged observations," which attempt to capture a dimension of classroom life that ordinarily remains undocumented—namely, children's talk.

FIGURE 1. Excerpts from *Butterfly Net*

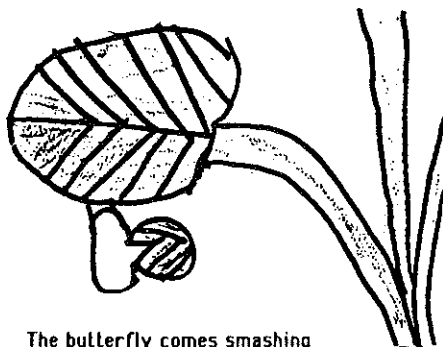
Four pages from a class book, compiled by the teacher, taken from children's discussions, writings, and drawings. Each page reflects contributions of several children.

It's biting its skin.
It's splitting its skin, sort of like a zipper.



It's like stepping out of your pajamas
with another set of clothes on underneath.

Its head is coming out
to grab on to the chrysalis.
If its back came out first it would maybe fall.
It looks like a space shuttle's door
when it's pushing open.



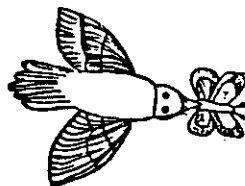
The butterfly comes smashing
out of the chrysalis.



Scientists' studies show
a butterfly has to pump up its wings
before it can fly.

Catbirds can eat Monarchs.

If other birds eat Monarchs, they will get
s'ck.



When a bird throws up from eating the
Monarch, it lets the bird learn his lesson.

Conducted according to guidelines, the discussions were tailored to purposes of assessment more than instruction, although the distinction is not meant to be a sharp one. The teachers held back somewhat from correcting or modifying children's comments; they also made deliberate efforts to involve most of the children. And, of course, steps had to be taken to make a record of the comments. The intent was to encourage broad participation, allowing children considerable leeway in what they chose to say or pursue.

Although the teachers' role in conducting discussions was tilted toward observer and listener, this is not to imply passivity. The teachers set the stage with initial questions, maintained the tempo, and reminded children of ground rules when necessary. On some occasions, teachers did intercede more directly, with the intent—if not always the effect—of steering the conversation in a somewhat different direction. And, toward the conclusion of a discussion, many teachers summarized or restated some of the ideas that had been raised, for the purpose of encouraging further thought and inquiry on the children's part.

Involving children on their own terms, within group discussion, requires skill on the teacher's part (Cazden, 1988). The teacher must support the interactions without dominating them and be willing to accept, at least for the time being, statements that are clearly "wrong" scientifically. But the patterns and themes in children's comments reveal much about their science knowledge and interests. The interactions among children may well bring out lines of thinking that would not come to the surface in individual interviews or group lessons, when children must deal more directly with the adult's agenda.

In primary science, the teacher's instructional decisions are often navigational in character. Decisions about curriculum and classroom program entail evaluation of the direction of children's work and thought, along with attention to details of their learning. Is a particular activity productive? Leading somewhere? Should attempts be made to alter the course of inquiry? Have the children reached a plateau? A dead end? Does something need refreshing? A new look? There are also questions of pace and timing. Some topics and activities need to be revisited over the year, given time for observation, or put on simmer.

For the teacher, discussions can be a time to take stock and set bearings. The focus is not so much upon individual learning as upon overall directions within the group. As we have attempted to illustrate, the value of discussions—for teacher or researcher—is that they provide a perspective upon the "woods" rather than the

“trees.” The patterns and themes in the children’s comments can guide educational decisions.

Qualities in Children’s Science Learning

In the earlier sections of this paper, we described specific findings to illustrate the outcomes of analysis of children’s discussions. Beyond those specific findings, the discussion data amplify several qualities of science learning that deserve attention for their implications for assessment and evaluation of learning. Three qualities are discussed in the following sections: dynamic nature of knowledge; realms of experience; and community of learning.

Background and Foreground Knowledge

Much of contemporary research on children’s ideas, intuitions, and concepts in science points to the importance of evaluating children’s background knowledge and prior experiences. Although disagreement exists about the specific instructional implications of children’s misconceptions or obstinate ideas (e.g., see Carey, 1986; Easley, 1984) there is consensus that we need to pay attention to the match or mismatch between educational programs and the child’s framework.

The discussions recorded in the present study provide another method for finding out about children’s background knowledge. Certainly, the children’s pronouncements and reported observations carried indications of their previous experiences and of their levels of understanding. Their reports of everyday experiences with shadows and reflection, and their interpretation of facts about the earth, their encounters with insects all provided examples.

But discussions are not simply occasions for displaying acquired information and ideas. Equally, they offer the teacher or researcher some indications of “where thought is tending.” What issues, questions, or problems seem to command children’s attention? Our analysis of the data showed that some substantive matters were pursued quite consistently by the groups, while others were not; the direction of children’s thinking was sometimes discernible. Some of the questions the teachers posed undoubtedly helped bring out evidence of this kind:

- What questions do you have about...?
- What have you wondered about...?
- What have you noticed about...?

The dynamic quality of the discussions reminds us that “background knowledge” is not a static collection of information.

Children formulate their ideas, opinions, and speculations in an effort to make sense of the complexity of the world. The facts that they do indeed collect, the ideas they talk about, the things they wonder about, are personally meaningful in some way. The movement of the earth, the appearance of the moon, shadows that change size, the fate of caterpillars are nontrivial and worth pondering. What we term "background" knowledge is inextricably linked to interest and question. And, one of the strengths of the discussion "method" as an assessment strategy is that it elucidates the connection between prior thought/experience and present focus. We, of course, need to know something about children's background knowledge, but equally we need to know what matters to them—foreground knowledge and the direction of their intellectual energies.

Realms of Experience

The interplay between prior experience and interests assumed different forms for each of the three topic areas investigated in this study. Correspondingly, each area posed distinctive challenges to instruction and, therefore, to assessment.

Of the three topics, the children's knowledge of insects seemed the most grounded in the resources and activities of the classroom. Much of what the children learned and thought about insects could be linked fairly directly to classroom opportunities: The bugs and the books introduced in the school setting were raw materials for much of their thinking. An assessment question that this situation raises is, What do children extrapolate from the study of life cycles and mini-ecosystems in the classroom? And, Does the experience of close observation generalize in some fashion?

The topic of shadow/light/reflection posed an instructional challenge of a different order. For this topic, all children presumed some ready-made expertise as they drew upon their everyday experiences. One of the instructional tasks in this case was to help children look beyond such experiences through investigations of light and shadow that would lead to productive questions. But the everydayness and obviousness of the phenomena under consideration made this difficult. As one teacher remarked, getting children to think about light was a little like trying to get fish to discuss the tides. How, for example, to move children toward consideration of light and light sources? How to build upon their natural interest in shadows' size, shape, movement—encouraging them to observe more systematically and critically?

Finally, the topics of earth/sun/moon in some ways raised the most troubling questions for teachers. Clearly, the children's

interests were strong; clearly, too, many of their ideas were a strange mixture of facts from secondary sources and, as one child put it, "facts from my head." Some of their questions were impossibly large. Compounding the problem is the ambiguity of key words: "earth," "moves," "round." These topics also do not lend themselves well to hands-on activities. For such reasons, teachers felt that opportunities for discussions were especially important. Children need the chance to "sort out" their ideas in a social context that allows expression of necessarily tentative understandings. The very process of conversation can help children examine ideas about complex matters that most of us—adult and child—will never fully understand.

The three topic areas were initially selected for the study because they represented the three broad disciplines that customarily frame the science content of elementary curricula: physical sciences, earth and space sciences, and life sciences. The children's discussions opened up these topics, indicating their diversity as domains of inquiry. Comparisons across topics suggested that each, from the young learner's vantage point, embodied quite different realms of experience and was associated with different modes of learning, different sources of information, and different points of access.

Social Quality of Children's Science Learning

In classrooms participating in this study, science activities are cooperative ventures—animals, plants, books, and equipment are shared. Children talk about what they are doing, informally in small groups and, occasionally, in larger meetings.

The children's discussions recorded in this study highlighted the communal quality of primary science learning. That children were capable of sustained interaction within the formal setting of a group meeting was one indication. As noted, such capacity undoubtedly speaks to the skill of the teachers and to the traditions of their classrooms, as well as to the developing abilities of children. As the teachers are quick to point out, coherent group discussions do not necessarily happen automatically. On the other hand, teachers were also convinced that children's strong interest in the topics did much to contribute to focus and cohesiveness.

A second indication of the social nature of science learning was evident in the children's shared observations and speculations. The children recognized each others' reports of experiences with shadow and reflection; their discussions provided evidence that certain issues about the earth's movement or the power of the sun were ones that many children talked about—theorized about—

outside of class discussions; and, in the insect discussions especially, their remarks provided strong evidence of shared observations.

CONCLUSION

The value of the study's methods for the participating teachers is worth noting. The process of staging and recording the discussions yielded documents that could be reviewed and considered as indicators of learning, and which served to complement other measures of growth. In primary grades, the evidence of children's learning customarily takes quite tangible forms of work products: artifacts such as drawings, writings, graphs, constructions with materials, and collections are in evidence. However, evidence of children's knowledge as manifest in discussion and conversation is rarely documented in such concrete forms.

The elusive or transparent quality of language means that teachers who value discussion for its contribution to children's thinking may nevertheless be left with the empty feeling that "we just talked," since few documents remain as evidence of learning. Children's work products abound, and can be reviewed and evaluated by teachers, shared with parents, collected over time; the children's conversations are less amenable to study. One by-product of this is that teachers become less certain about what is happening when discussions are held. Are the discussions really productive? Contributing to thought? To development of ideas? Or, as one child remarked toward the conclusion of a discussion, "Can we do science now?"

The need for assessment approaches that tap children's conversations is clear. Aside from benefits for teachers, such approaches move toward an assessment model that validates the cooperative, interactive characteristics of young children's science. Methods of assessment should be congruent with principles of practice. In the long run, the design of assessment strategies has consequences far beyond the particular data that might be obtained.

Children's Investigations of Natural Phenomena: A Source of Data for Assessment in Elementary School Science

Hubert M. Dyasi

A VIEW OF ELEMENTARY SCHOOL SCIENCE

At the City College Workshop Center in New York, we regard elementary science as encompassing content and approach, not merely scientific generalizations. The content is the common materials and phenomena we all encounter at one time or another; the approach is inquiry built around observations and experience and around making meaning out of commonly occurring phenomena. Direct experience with phenomena in the world connects the content of elementary science with children's experiences and observations outside the classroom. The connection is not trivial. On the contrary, it reconfirms that science is a continuing search for underlying commonalities in apparently disparate phenomena and an intense engagement with things that arouse curiosity in us.

In an illustration of this approach, Jos Elsgest, a Dutch science educator who worked for many years in science education programs in Africa, engaged African children in a study of the ant lion. The children's curiosity had led them to wonder about observable characteristics of the ant lion and, assisted by skillful guidance from the teacher, to devise ways to answer their own questions through direct observation of the ant lion. A transcript of part of the teacher's record of one of the classroom discussions held after several lessons involving children's observations of the ant lion gives a flavor of the approach:

- T: Do you remember what you have already learnt about the ant lion [from the ant lion itself]?
- C1: They live in the soil.
- C2: They move backwards.
- C3: They like the sand.
- C4: They cannot live outside the sand.
- T: How do you know?
- C4: I tried it. I put it in my tin without sand, and it died.
- T: After how long did it die?
- C4: After three days.

Hubert M. Dyasi is Professor of Science Education at City University of New York and Director of the City College Workshop Center, where he educates teachers in making classrooms better contexts for active learning.

T: Why do you think it died outside the sand?

C4: It cannot live outside the sand.

C5: It could not eat.

T: Now, that is a big problem: What do ant lions eat?

As can be judged from this exchange, the children learned directly from the ant lion about where it lives and about its locomotion. The record went on to show that, among other things, they had also learned about what it eats, how it catches its prey, whether it can see, and how many legs it has.

The majority of today's elementary school classrooms are not conducive to that kind of elementary science. In schools, elementary science suffers from an overemphasis on the study of symbols (written words, verbalism, and rote learning), from a view that elementary science is secondary science made simple (that is, an academic emphasis), and from a belief that children's work on nature selected and pursued by children with little intervention by adults is aimless activity. Further impeding practical elementary school science instruction is the fact that elementary school schedules and classrooms are not organized for easy implementation of science inquiry. They do not create opportunities for first-hand familiarity with a variety of biological, physical, and man-made phenomena in the world around the child; they do not cultivate the child's interest in further, self-initiated exploration of that world; they do not provide opportunities for a child to develop and demonstrate knowledge.

Despite these impediments, many arguments favor an activity-based approach to elementary school science. Some of these arguments arise from research evidence and others from recommendations of prominent scientists. Among research studies that provide evidence that support this approach are those by Champagne (1988) and others cited in reviews by Bredderman (1983) and Shymansky, et al. (1983). The research shows that specific science education benefits accrue from activity-based science learning. These include higher mean scores in achievement, in perceptions, and in process and analytical skills for children in activity-based programs compared to children who learned elementary school science through other approaches.

Reflections of prominent experienced scientists and of keen observers of the scientific enterprise point to important resemblances between our view of elementary science as precursor to quality science inquiry and the practice of science. Philip and Phylis Morrison (1984) have said quite simply: "You can't talk about science and remain solely in the domain of symbolic discourse. You require some contact with that substance of which

science is a symbolic representation" (p.4). Arons (1983) expressed the point in these words:

Experience makes it increasingly clear that verbal presentations—lecturing to large groups of intellectually passive students and having them read text material—leave virtually nothing in the student's mind that is permanent or significant. Much less do they help the student attain what I consider the marks of a significantly literate person. (p.92)

Hawkins (1983) put the issue as follows: "There is a marvellous continuity between the worlds of children's experience and the adult worlds of the arts, of the sciences and mathematics, of conduct and social life...This continuity is one of cumulative learning." (p.65)

The connection between scientific and general cultural ideas and how both are accessible to us from the world we experience is absolutely important in elementary science because it underlines the concept of continuity between childhood and adulthood worlds as they relate to science learning. The concept of continuity of experiences and understanding is often overlooked in elementary school science teaching in favor of presenting science merely as an academic subject consisting of isolated facts, conclusions, technical words, and scientific laws. And, as a consequence, children tend to see science as related to neither their experiences with the world nor their conceptions of relationships among things. Yet children bring to the learning situation an enormous wealth of experiences, interpretations, and logical connections with the world upon which science learning can be built.

Children, just like adults, achieve a better understanding of the world by continually building and reinterpreting their direct knowledge, thus improving the conceptual frameworks they bring to their learning. Almost 35 years ago, Navarra (1955) conducted a study that demonstrated the validity of this association. From a vast number of close observations of a preschool child covering a period of over 12 months, he perceived a "persistence of growth and refinement in the child's experience" with phenomena of nature. The child continually made observations of events in the world around him, and the observations in turn led the child to form logical relationships among, and interpretations of, events. Navarra also found evidence that "the logical relationships were continually revised as further observation expanded the matrix of experiences." (p.35) Conceptual imperfection, change, and refinement on the part of children must be recognized as ingredients in elementary school science practice; they must be recognized because children's conceptions of the world undergo change as

children grow and gain more direct experience with the physical world and with the worlds of symbols and ideas. This notion of continual refinement and accretion of knowledge is associated closely with the development of scientific knowledge.

THE INTENTION OF SCIENCE AND CHILDREN'S WORK

In the Workshop Center approach we expect elementary school science learning activities to encompass three closely related aspects: primary inquiry into a phenomenon of nature, symbolic representation of one's observations of the phenomenon, and seeing patterns in the representations. Primary inquiry involves direct first-hand experiences with a selected piece of the natural world. If pendulums were selected, children would have direct experiences with pendulums—making them, examining them, and thus getting to know the parts of the pendulum; observing pendulum motion and how it varies under different configurations; and identifying periodic motion in nonclassical pendulums (e.g., motion of the arms of a runner).

If the subject was living things, children would interact directly with living things, noticing their characteristics, their habits and their habitats, changes that they undergo over time, and many other aspects that capture children's interests. Observations children make may initially be of a general nature, but will often be refined as a result of a teacher's quest for greater specificity. For example, children will move from describing something as an insect with wings and legs to an insect that has so many pairs of wings, with a specific shape, patterns of colors, venation, and specific points of attachment to the insect's body. They would similarly move toward more detailed observations of legs—e.g., jointedness, number of pairs, smooth or rough, soft or hard, and so on. The overall significance of primary inquiry is that children devise valid and reliable ways of obtaining, and do obtain, information directly from nature. Furthermore, they learn to raise and answer first-, second-, and even third-order questions while using materials.

Studies carried out in England by the Assessment of Performance Unit (APU) and reported by Wynne Harlen of the University of Liverpool (Harlen, 1985) showed that although most children's observations focused on gross features, children responded capably to seeing details when asked to do so. The studies also demonstrated that children's observations not only bear relevance to a specific task but also involve senses besides sight.

The second part of the process, representation of observations, takes a variety of forms—words, drawings, sound

recordings, numbers, tracings, and photos, for example. Regardless of the mechanisms of representation used, children learn that accuracy, relevance, and comprehensiveness are of central importance in representation. Representation of some aspect of nature in symbolic form encourages closer observation of detail for purposes of keeping an accurate record and of communicating observations. Accurate portrayal of what was observed means close observation of the selected phenomenon and use of tools to refine the observation; for a description to be relevant, children must select for observation largely those items deemed important. Comprehensiveness refers to the breadth of information gathered about the relevant characteristics.

When children detect patterns as a result of their observations and representations of nature, they are engaged in generating knowledge and in developing concepts. At an ordinary level, the activity might involve seeing similarities among different things. At a slightly deeper level, it involves inventing categories of attributes shared by objects or organisms. At a very deep intellectual level, children discern a pattern in the attributes of the same thing under varying conditions or configurations. When it is not trial and error, the process involves studying the descriptions, manipulating variables and conducting tests to yield more descriptions, creating an organizing scheme to establish order from the descriptions, and drawing conclusions or abstractions based on evidence. The abstractions resulting from this process are examples of going beyond lists of observations and their representations to careful operational statements that provide a basis for making predictions and for developing additional understandings. And that is the intention of science.

There is a difference, of course, and a big one at that, between scientists' science and children's science. The difference resides at least in the frames of reference and ideas of children on the one hand and of scientists on the other. A classroom example in which children were allowed to reveal their understandings shows the relationship between observation and a frame of reference. Hein (1968) followed the work of fifth graders studying linear motion of objects of different shapes down an inclined plane. The children had been asked to compare ways the different objects moved down the inclined plane; and they did, but not in ways the science educators expected. The children raced the objects against one another to see who the winners and who the losers were. No matter what questions the educators asked or in what direction they tried to lead the children, the children persisted in looking at the events as races. Within the frame of reference of

“races,” children made some observations and failed to take note of other events that would be important in a different frame of reference. In the children’s frame of reference, the important observation was spotting the winners; ties were irrelevant. Hein concluded logically and reasonably that “these children do not have a statistical view of data and scientific observation. Instead they have a particulate view of events. Each observation has its independent existence, each observation could decide the contest.” Looking for winners in this activity is not what a scientist would do. “Ties” would also be very important in a scientist’s frame of reference of probability of occurrence of independent events.

A second difference is that children’s science tends to draw understandings directly from the nonidealized conditions we all know, whereas scientists’ views relate to established canons of knowledge drawn from idealized (or controlled laboratory) conditions. For example, children know that, in free fall, heavy objects fall faster than lighter ones. Scientists make the same observations, but children’s explanations of this phenomenon will differ from the scientists’ because children’s frame of reference in this case is centered only on the weight of the objects and does not encompass observations of free fall in a vacuum. The different frames of reference or presuppositions with which children and scientists approach this observation result in different “facts,” different trials of the effect of one factor on another, and different degrees of elaborateness of investigations. In her paper on primary school science, Lillian Weber, founder of the City College Workshop Center, discussed this point in detail, drawing attention especially to the fact that “a child’s inquiry takes place in a world of people and things” rather than in the well-ordered laboratory environment of the scientist (Weber, 1973). “Long before and even after he has acquired words,” she went on, “his [the child’s] asking can be observed in his actions, and his tentative solutions in his adaptations of his actions. His actions, in fact, are the equivalent of hypotheses.”

Apart from the curiosity that underlies both children’s and scientists’ science investigations, other very important resemblances between children’s science and scientists’ must not be overlooked. In the example cited above, Hein and his colleagues reported that children approached problem situations from a preconceived frame of reference (in their case the framework was competition) that provided a basis for making decisions about which attributes were essential for making judgments about events or for making sense of the observations. Scientists also look at and interpret observations from the context of a frame of reference.

This is a basic similarity between child and scientist. If the frame of reference of winners and losers is flawed as a basis for making scientific understandings in the case of rolling spheres and cylinders down inclined planes, it has to be improved by stimulating children's interest in examining what still needs to be explained: the ties. Such an examination might lead children to consider frames of reference that allow for a more comprehensive and reliable description of observed events, to go, as it were, beyond "naive" notions or theories to careful operational statements that lay ground for predictions and for broader understandings. This progress toward authentic scientific action and thought should be the focus of assessment. A close look at children's work suggests that the work can be an important point of focus in the assessment of children's progress in the learning of science inquiry.

EXAMPLES OF CHILDREN'S WORK

When children in classrooms are allowed to carry out primary observations of nature in their environment, to construct representations of their observations, and to detect patterns, science inquiry thrives. A fifth-grade class that had been looking at insects for several lessons developed its own classification scheme (table 1) and a "key" for identifying insects found locally.

A person looking at the children's classification scheme might be struck by its unusual basis and by errors it contains. For example, a person might think that the first column is not necessary—that the categories are also characteristics in some sense. One notices also that some of the organisms belong to more than one "class." From the Linnean frame of classification (the genus and species frame of reference), earthworms should not be included because they are not insects. These are legitimate sources of concern, but the concern must not overshadow the power of the children's creation of a scheme.

Other children in an elementary school class in England engaged in a similar science learning activity. One of them developed the classification scheme shown in table 2 (Rowland, 1984).

The teacher reported that in this classification activity, the teacher reported that the child *first thought* about the attributes he wanted to use and then examined the specimens over and over again and selected those that shared the attribute. This thinking about an embracing attribute from discrete observations is a very bold and constructive intellectual activity whether it is done at the frontiers of a discipline or, as in this case, in earlier stages of learning. The action signifies the interpretation of nature on the

TABLE 1. Children's Scheme for Classifying Insects

Category	Characteristic	Examples
live communally	very small, build houses	wasp white ant.
attack intruders	eat other insects or powdery things	bee, safari ant
Do not bite	humble earthworm	mantids, butterfly ant lion
Make honey	lay eggs, obtain food from flowers, suck food, hairy legs	bee, butterfly,
Have three parts	short wings, eat tender leaves, big eyes, winged	wasp, house fly, termite, bee
Live underground	no eyes	ant lion, white ant, small black ant, safari ant
Live in trees	eat fruits, bite	white ant

TABLE 2. A Child's Classification of Caterpillars

Colour	Fatness	Hairy	Found on	Sameness
green yellow and	thin	not hairy	hawthorn	not the same
grey brown	fat fat	bit hairy hairy	hawthorn dock leaf	not the same
brown	medium	bit hairy	hawthorn	same

basis of observations, representations, and understanding of the selected organisms in the environment.

The creation of the interpretive scheme shows that the children have gone beyond particular examples to think of generalizations that can be supported by demonstrable observations. To fail in the classroom to build upon the natural inclination to make connections and to create schemes that account for perceived relationships (among living organisms in this case), and to fail to capture the same in assessment, would be to miss an important aspect of children's growth in science inquiry. What remains to be done to further the science development of the elementary school children whose work is referred to above is no small task: It is to encourage them to be willing to refine and modify their scheme as finer distinctions need to be made and to help them learn how to do so. Before children could evaluate the usefulness of this scheme, they would have to use it extensively. From that use perhaps they would recognize problems associated with a scheme that does not, for example, discriminate well among things that are very different from one another in some important respects. In time, they might see the value of seeking guidance from schemes developed by others. Perhaps as they observe other organisms closely, their attention will be drawn productively to looking at the structural characteristics of organisms in order to make fine classification distinctions, and they will recognize their earlier schemes as first approximations that were useful for gaining a general idea and for laying a foundation for a coherent picture.

There is power in good observation, a sense that one knows and that one sees connections. And observations need not be represented only in prose and drawings; they can be represented in verse as well, as 10-year-old Leo's poem shows:

It's a Spider
Moving through the night
As if always in flight
From some unseen enemy.
In the summer webs on trees
In the fall webs in the leaves
In the winter you die on out
In the spring your children
Search for a new home.

(Leo—not his real name—did this work at the Prospect School, Bennington, Vermont).

There is a sense here that Leo has focused on the spider not momentarily, but over an expanse of time and of space and has

arrived at the notion of the physical home (the web) located in a broader habitat (the tree at one time and the leaf at another). Leo captures the spider's life cycle—life of the spider in the summer and fall; death in the winter; and then the young ones appear in the spring. Unstated but understood is that in the summer these young spiders become adults who, presumably, will die the next winter. The young ones have the task of building a home or searching for one. The great explosion of life in the spring and the end of a lifetime in the winter have been duly noticed and recorded. *That* is the essence of observation to make meaning (Carini, 1979).

The examples of children's work given in this paper indicate important points about children and their learning about nature. Children can generate knowledge directly from objects in nature, and such knowledge has characteristics beyond those of mere speculation and guessing. These characteristics include obtaining information from direct experience with concrete natural phenomena through systematic manipulation and observation, utilizing symbolic material to represent the observations faithfully, and making relevant abstractions from the representations. The challenge for assessment is to find strategies and mechanisms that portray children's developing characteristics in these aspects of elementary school science learning.

DOCUMENTATION OF CHILDREN'S WORK

Documentation of children's work over *significantly long periods of time* is one of the best sources for assessing children's progress in science inquiry. The documentation can be obtained through the research technique of intensive observation and recording of a single child's experiences and responses. As indicated above, Navarra (1955) used this method to study the development of scientific concepts in one child. Although this method yields invaluable information for assessment purposes, it cannot be used consistently by a classroom teacher who has, of necessity, responsibility for all the children in his/her class and for only nine months. But it can be modified to meet the constraints within which the classroom teacher interacts with children. The records of children's work at The Prospect Archive and Center for Education and Research in North Bennington, Vermont, are an excellent example of such a modification. The Prospect Archive is a unique collection of individual children's drawings, writings, constructions, and other artifacts spanning an average of six-to-eight years of a child's school experience. The material on each child also includes teacher's statements about the child's educational activities on a week-by-week basis and a general summary covering

each term. Below is an excerpt from a teacher's general summary about Leo, the child whose work was cited above (The Prospect Archive, 1984):

He (Leo) builds intricate structures all of which have long explanations to go with them. One building of (Leo's) was... a building on another planet complete with laboratory, energy sources, water systems, solar collectors, secret passageways with trap doors. (Leo) has a natural sense of balance and symmetry... He is very inventive with wood and thinks up very original projects for himself to do. He built a base for a star ship. For this he invented a pivotal cannon that could move up and down and around. It was very impressive because he had come up with the whole thing completely independently. (p.54)

These evaluative statements are part of the data attached to the child's work. Interested people, such as assessors, can have access to the entire portfolio to make their own judgments. The teacher's statements do not make reference to the inquiry process associated with these activities, but there could have been such references had the teacher included inquiry learning as a major focus of the child's activities. But the teacher did view making representations of objects as a very important activity for the child, hence the following comments:

(Leo's) drawings often express his mechanical interests. They are often cross sections of buildings revealing all the inner networks of stairways, water systems, energy systems, and structural supports. His drawings are striking for the detail and depth. (*loc. cit.*)

The teacher's comments indicate quite vividly what a close observer Leo is; they lead one to yearn for a direct look at the child's work to satisfy one's curiosity about it.

A third mechanism is the documentation of group activities within a class over extended periods in the form of a teacher's journal. In this case, work of groups of children is accumulated over time, thus creating a bank of detailed material encompassing aspects of science learning activities. The children's work cited above can be part of a bank. Assessors can use material from such a bank to make inferences about the children's progress in science learning. Included with the children's work would be their teacher's perceptions and reflections about the work. *Juba Beach*, a teacher's journal prepared for, and published by, the African Primary Science Program, is an example of such a journal. The journal includes children's descriptions of their science inquiry learning activities complete with written accounts, diagrams, and

questions related to the organisms the children studied along a beach. The teacher's comments, interspersed in the children's own accounts, are informative. For example, in Juba Beach the teacher wrote:

The general topic of beaches and sea integrated many experiences of learning. The children found and observed a wide variety of animals. They examined rocks and shells and sand. They tasted and tested water for salt content. They counted waves and the flow of rivers and talked to fishermen. The challenges were without limit... The events of this unit encouraged them to find answers to new questions. They wanted to learn and because of this they used and developed their skills—they measured, weighed, compared and counted, they kept notes and discussed their findings. For me, their own evaluations and this record book tell more about the progress of the children than any written examination I might have given them." (*Juba Beach*, A Unit of the African Primary Science Program. Education Development Center, Newton, MA, 1971)

Computer technology can be used to build a bank of data on the work of children in middle- and upper-elementary school grades, recording and storing evidence of the quality of their participation in science inquiry learning activities. In such cases, the computer is a tool that children use to record their observations, experiments, and abstractions derived from their science investigations. The records can be retrieved by the children, the teacher, or someone interested in them. The Bank Street College of Education's Center for Children and Technology in New York City has done interesting work in this respect. In the *Earth Lab* project, children work in groups to do earth science inquiries; they collect data and later share their findings. The Center's software project, *INQUIRE*, enables children to record their ideas; keep notes; record their plans, their guesses of expected findings, and their findings while engaged in inquiry activities on sports physics. As a result, children create their portfolios as they progress in their elementary school science learning activities during science investigations.

Another interesting use of computer technology in elementary school science learning that has great potential as a documenting mechanism for assessment purposes is telecommunications used in The National Geographic Kids Network project. This project for grades 4–6 is carried out jointly with Technical Education Research Centers and combines the use of computers with telecommunications. Children in the network conduct experiments

in their local areas, collecting data on acid rain, for example. They link with children in other localities by sending the results of their local experiments through the telecommunications network to a central computer. Through the network, children in various parts of the world can discuss their findings with their peers and work collaboratively in a manner similar to how a research team works. Although many classrooms might have difficulty gaining access to a telephone line, the computer component of the activities can be good for record keeping.

One of the most common components of assessment is testing. Tests provide a snapshot of a student's performance in selected categories of science education. They may be paper and pencil and/or may involve students in carrying out practical investigations. The Assessment of Performance Unit of the Department of Education and Science (United Kingdom) has conducted a series of surveys to assess science performance by elementary school children, specifically, to assess "how well children can use the process skills of science and apply ideas in solving problems through investigation." In 1982, the APU conducted a science performance survey of 11-year-olds (DES, 1984). This survey's practical test assessed children's ability to plan investigations. It was administered to children on an individual basis, i.e., each child was required to answer the questions on his/her own. Each of the four investigation problems presented to the children was scored on the following components: general approach (e.g., number of tests carried out), specifying use of equipment, varying the factor whose effect is being investigated, controlling variables, looking for the effect of changing the independent variable, repeating measurements, and recording and interpreting results (Department of Education and Science, 1984). These practical tests used no multiple-choice questions. The problems centered on common situations such as planning investigations to decide which of a set of wood blocks would make the best chopping board, what one would do to find out which paper towel holds more water, and so on. This type of test holds promise because it lets the child demonstrate all three aspects of doing quality science inquiry: focusing on the concrete material directly; making observations; varying and controlling variables; and drawing conclusions in a free-response situation. (A New York State Elementary Science Program Evaluation Test for fourth grade also includes a similar practical component.)

Within four years after the results of the UK surveys were published, another report of recommendations for assessment in

the United Kingdom appeared (DES, 1988). The report suggested attainment targets in precollege science at ages 7, 11, 14, and 16 years. The report identified what should be expected of pupils in their development in science at each of these ages and proposed that children be assessed against these expectations. It is too early to assess the impact of the proposals.

COMPARABILITY OF DOCUMENTATIONS

The view of elementary school science described at the beginning of this paper provides a context for assessment practices focusing on children's work collected over long periods of time, such as one academic year. Specifics within that context include children's questions and their efforts to answer these questions through direct study of phenomena, as well as children's ability to design experiments that show manipulation of variables and testing of "hypotheses," to make sense of collected data, to communicate information, and to raise further questions that yield more data. These can serve as foci for assessing the science education quality of the work. The question is: Who does the assessment?

Even before answering that question, however, we need to ask another: Could the kind of documentation of children's work described in this paper be done in ordinary elementary schools? The answer is that most of it can be done in ordinary classrooms by suitably educated teachers enjoying unfettered professional judgments. The few examples of such teachers described in this paper indicate the possibility. Before inservice as well as prospective teachers can engage in this science education approach and documentation, constraints placed upon them in schools and in teacher development programs would have to be adjusted.

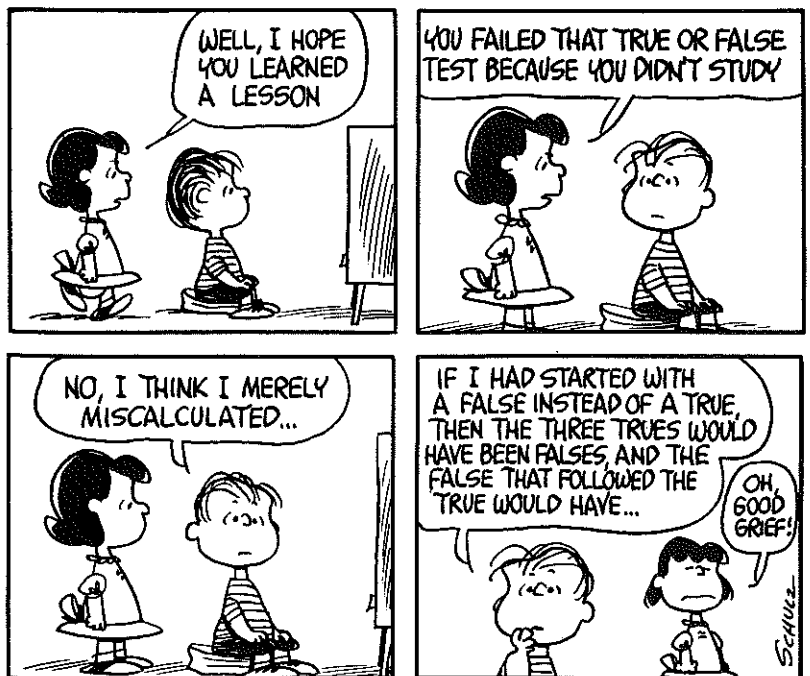
Teachers would be the primary people assessing children's growth in elementary school science. They would use prepared assessment guidelines indicating how the children's work is to be judged. Preparing the guidelines would be one of the most important activities—the guidelines would be sufficiently flexible, unambiguous, and faithful to the advocated science education approach. They would be prepared by carefully selected panels of teachers, science educators, appropriate child-development specialists, leading scientists, and school children. Since community schools are local institutions, groups of teachers at the local level and assessment specialists, science educators, and, if possible, outstanding scientists, would come together to examine the children's work. From that examination, they would prepare detailed reports describing the work, indicating how it was assessed, and

making illustrated evaluative accounts. Local school districts or the state would decide whether to make assessments by child, grade, school, district, or by a combination of the above. For purposes of comparison across school districts, panels of teachers and assessment specialists from the districts, as well as members from other states if desired, would examine samples of assessed children's work. These panels would also prepare detailed, professional reports that could be used widely to improve elementary school science instruction. From this start, improved ways of assessing children's work could develop.

CONCLUDING REMARKS

Assessment in general, and in elementary school science education in particular, does not occur in a vacuum. It is embedded within a preferred educational approach and serves clear purposes. In this paper I have portrayed elementary science education as focusing on children's engagement in organized inquiry into their world. I have also implied that learning science inquiry implies a considerable density of experience with things of nature over long periods of time. I view one major purpose of assessment as support for quality instruction for the inquiry approach to science outlined at the beginning of this paper. The examples of children's work indicate that children in ordinary elementary schools are capable of conducting inquiries into nature, but in their own idiosyncratic ways. It is through an examination of children's work that one gains insight into their stage of development in the attainment of the art, skill, and knowledge of doing science. An assessment that focuses on the various facets of that work in an integrated fashion would reinforce classrooms that encourage children's inquiry activities. I am calling, therefore, for assessment that portrays children's stages of development in science inquiry in appropriate practical, concrete, and investigative activities; an assessment that also portrays children's capability to communicate their stage of development through the questions they raise about nature, the observations they choose to make, the symbols (such as written words, numbers, drawings, tables, graphs) they use, and any other ways they devise. We have ways of documenting such data; what we need is willingness to implement. That willingness will not come about until educators and assessors accept that elementary school children's inquiries are valuable starting and continuing points for science inquiry instruction and that elementary school teachers can play a large role in elementary school science assessment activities.

CONCLUSION



Conclusion

George E. Hein

INTRODUCTION

The papers in this volume all discuss issues concerning appropriate assessment for hands-on science programs. Each author has provided material on a specific topic, but these papers also have commonalities. In this concluding chapter, I will summarize some of the issues that repeatedly emerge from these valuable contributions. This chapter has been immensely enriched by the discussion during the three days of the conference. I have repeatedly listened to the tapes of the meeting, and each time I am impressed by the wealth of topics, the warmth of the discussion, and the clarity of the ideas presented. Much of what is written here I owe to the contributions made not only by those who prepared papers but also by the other participants. I have attempted to represent views fully and combine ideas to provide a coherent whole for the readers of this volume. I apologize if I have slighted anyone's contribution or not described it accurately. Although I don't and can't take credit for all the ideas in this chapter, I do take full responsibility for discussing them.

WHAT IS ELEMENTARY SCIENCE?

The first task of any assessment effort is to define the activity or domain to be assessed. If we want to know how well someone reads, plays the violin, or sells real estate, we have to define what we mean by those activities. Although this requirement is obvious, how to describe the content of elementary school science is not clear. This issue, of course, underlies any discussion of validity. An assessment is valid only if it matches that which is to be assessed.

What do we expect a child to be doing when that child "does" science? What would be the ideal evidence to use to assess science competence? Surprising but true, there is no commonly accepted answer to this question. In contrast, if we ask what the ideal reading assessment should determine, most people would agree that a reader should be able to pick up written material, read it, and comprehend the content—that is, derive meaning from the written text. There is even general agreement about the types of books that elementary students should be able to comprehend at

various stages of development. Educators may disagree about *how* comprehension should be assessed: Does a reasonable test of that comprehension require a reader to perform some physical act (e.g., follow the written instructions)? Does it require that the reader retell the story? That the reader answer closed or open-ended questions? But all these issues are usually discussed within the framework of a general agreement on what it means to read with comprehension. We can make no comparable statement for scientific competence, especially at the elementary level. We have not agreed on what it means to “do” elementary science with comprehension. Even the framework of the discussion may vary significantly. Does science competence consist of knowledge of facts and concepts, or is science a process, manifested only through practical accomplishments? Should there be a canonical set of information and/or activities that all elementary school children master, or can the content vary from school to school? Should the goal of science education be to produce scientifically literate citizens or to prepare students to become scientists?

If we attempt to answer this question by analyzing the material currently taught in U.S. elementary schools, the answer will fall within two general spheres: Either elementary science consists of a set of facts about and descriptions of the natural world, and/or elementary science consists of a set of rather noble but vague goals about attitudes toward the natural world. Neither of these provides an adequate basis for developing a comprehensive assessment scheme. The former definition can lead only to assessments that focus on facts, while the latter provides little guidance for any serious attempt at assessment.

Some recent descriptions of what science instruction should encompass include both content and the processes of science. They also include statements concerning attitudes about science and values. But the translation of these statements into actual curriculum is not simple, and the development of a canonical body of knowledge may not be appropriate. For example, the AAAS report *Science for All Americans* (Project 2061) has provided an impressive description of desired outcomes in science. But it still lacks specificity. Are there activities that it recommends for all students? Are there explorations of the world to which every child should be exposed?

At the conference, participants discussed the relative importance of process and content. The following issues appear to me to be most important in this discussion:

a) Arguments concerning the relationship between process and content need to be developed with the recognition that neither

exists in isolation. There can be no significant process without context, and there can be no important content without the processes of science. It is impossible to learn to measure in the abstract—measurement must be of something. Observation skills exist only when they are applied to some event or object. In her paper on the APU assessment scheme, Patricia Murphy describes one approach that uses specific content to assess students' grasp of the processes of science. The categories of the assessment framework all refer to processes (using symbolic representation, observation, planning investigations, etc.), but the actual questions all refer to specific tasks or concepts: How would you find out whether wood lice prefer light or dark? Give two reasons why moss grows only on three sides of the pictured tree trunks, etc. In contrast, U.S. assessment systems often include process or theory questions asked outside of a specific context. See, for example, item 1-b from the Connecticut assessment quoted in Joan Baron's paper.

b) Scientific content actually includes two qualitatively different components: the concepts of science and the underlying body of information that contributes to developing these concepts. The importance of concepts, and their development over time, is stressed in Rosalind Driver's paper, and referred to in several others. A growing body of research literature documents that children's concepts may be very different from current scientifically accepted views. Children (and many adults) hold Aristotelian views about the laws of motion, properties of light, and so forth. These views are difficult to change, especially since they are often supported by considerable common-sense evidence. Therefore, one part of the content of science involves teaching concepts.

But another component of the content of science is the direct content, the facts of science: Water freezes at 0°C.; tadpoles turn into frogs; heavy objects fall at the same rate as light objects. Unfortunately, the relation between facts and concepts is itself not simple, and there are many examples in science (Kuhn, 1970) and in pedagogy of cases where the facts are altered to fit the current theory. For example, Hein (1968) noted that fifth-grade students consistently "observed" that when two tubes were rolled down an inclined plane one always "won" the "race," even if they were identical. Adult observers had a different theoretical framework. They knew that identical tubes reached the end point simultaneously and observed that the time the tubes took to reach the end point was equal.

A comprehensive statement describing the content of elementary school science needs to consider both concepts and facts and how they relate to each other.

c) Especially in elementary science assessment, we must define what we mean by scientific knowledge at the level of six- to 12-year-old children. Is this the same as adult scientific knowledge? Although the technical ability to read is fairly well established by the end of the elementary school years, even an ideal reading assessment would have to be adjusted for the ages of the students involved. If a third-grade student failed to comprehend Wittgenstein's *Philosophical Investigations*, we would hardly be justified in concluding that the child could not read. What are the similar levels of scientific understanding that we can expect from a nine-year-old child? Is it appropriate to expect a child that age to understand that, contrary to the evidence of the senses, the earth moves around the sun? After all, the best human minds, applying their rather impressive intellectual abilities to all that they could observe, believed the opposite for millennia. Can we expect a nine-year-old to comprehend both that scientific statements should only be accepted if they are supported directly by evidence (or as it is sometimes described in science texts, that nothing should be taken on authority), and that matter is composed of atoms, but these cannot be seen or felt, and statements by scientists concerning their existence should be accepted based on the indirect evidence available?

Formal Science and Everyday Life

Another issue concerns the relationship between "formal" science and everyday life. Traditionally, school science has focused on activities and concepts that are important to professional science, often material that is significant in the history of science. When I first became interested in elementary science instruction in the 1960s, most elementary science texts advocated that teachers heat mercuric oxide in a test tube to demonstrate evidence for the atomic theory. Heating mercuric oxide changes an orange powder in the bottom of a test tube into a silvery film of mercury on the sides of the tube. It also requires an open flame in a classroom and releases highly toxic mercury vapor into the atmosphere. Why would anyone ever have recommended this experiment? In part, our environmental and safety awareness was less sophisticated then than it is now. But, more important, this experiment happens to be of great historic significance; it was a crucial one for Lavoisier in his discovery of oxygen and his experiments that challenged the phlogiston theory. Although there are other simple and safe ways to demonstrate chemical transformations, the experiment remained in the texts as an homage to tradition. A similar situation would exist in reading instruction if we required children to peruse McGuffey's readers and expected these moralistic texts from another era to speak to and provide meaning for them.

Another component of elementary school science that I believe has more historical significance than actual value is the use of pendulums in classrooms to demonstrate simple laws of mechanics. Pendulums are important in the history of science. They played a key role in Galileo's reformulation of physical laws. This curious device interested the restless, imaginative Italian scientific genius, and he noted that the period of pendulum swings was independent of the weight of the pendulum bob. That conclusion, which could be easily duplicated by others, was one of a number (including the famous experiments of dropping objects from a height) that led Galileo to his ideas about motion.

It is a long leap from this bit of history to the assumption that experiments on the periodicity of pendulums *should* be part of an elementary science curriculum. When pendulums are introduced into the classroom, children frequently carry out all sorts of exciting activities with this unfamiliar object, but they don't spontaneously do the "right" experiments. An excellent example of a classroom in which the teacher subtly and not so subtly needs to direct the children to the appropriate experiments on pendulums is provided in Edwards and Mercer's (1987) ethnographic study.

School Learning and Nonschool Learning

Brenda Engel discusses the distinction between school learning and before-school and out-of-school learning. Unfortunately, children are usually required to consider the reading and prereading tasks they carry out at home as distinct from the formal reading instruction at school. Similarly, we need to be clear about the relationship between the everyday science human beings know and use to survive in the world and what is taught in school as science. Most human beings can reason in some domain, devise rough-and-ready experiments, draw inferences, and observe consequences in some sphere. Children learn to play strategic games, to build with blocks, clay or wood, to garden or to cook. Some are interested in nature, others in mechanical objects. These interests in exploring the world are the natural equivalent of trying to find meaning in written text; they represent prescience and early science activities as much as do spontaneous efforts to read and write.

To what extent should elementary science include the everyday activities of life and to what extent should it include the specialized experiments, data, and activities associated with formal science? The APU experience suggests that both are important and that we may achieve different assessment results depending on which we stress. For example, problems set in terms of scientific equipment and scientific language may produce quite different responses from similar problems set in everyday language and

using kitchen or other household equipment as illustrations. Students' use of qualitative or quantitative strategies, gender differences in responses, and ways in which students perceive the problems may all vary (Gott and Murphy, 1987).

LEVELS OF ASSESSMENT

A common theme of many of the papers is the wide range of activities that can be considered as part of assessment. In her discussion of literacy assessment, Engel cites a multitude of methods that can be used to track students' progress; in the APU surveys, Murphy describes how various categories of student achievement are assessed through several kinds of written tests as well as individual and group practical tasks; the final two papers in the volume describe the use of classroom discussions and collections of children's work as assessment methods.

We need to acknowledge the range of possibilities for assessment. It is worthwhile to relate them along two different continua. One continuum is the range of *methods* possible for assessment. These methods can be classified in three major groupings.

a) At one end of the continuum is any system used to keep track of what happens. Whether this *record keeping* consists of informal notes by a teacher, a folder of student products, or check marks on a list of milestones, record keeping usually takes place at the level of the classroom and involves only the people actually engaged in the activity.

b) *Documentation*, any systematic manner of preserving documents for assessment purposes, is the next level of assessment methods. These documents may consist of collections of products or notes or materials collected specifically for assessment purposes. Periodic interviews with children, recorded conversations about science, key samples of tasks completed, portfolios, or even collections of test results may be included in documentation.

c) *Testing* represents the most formal level of assessment in that, by definition, the activity used as a test must be specifically designed to serve that purpose. This does not mean, of course, that a test needs to be a multiple-choice, short-answer activity. A wide array of performance measures can be used as tests, from a variety of paper-and-pencil methods (Carlson, 1985), including essays, to performances, trials, auditions, etc. But these all have one thing in common: They are specifically carried out as assessment tasks, and a scheme for rating the performance on the activity has been agreed upon. Tests may even be embedded within the day-to-day activities of elementary science instruction (or any other activity) so that the participants—the children in the classroom—are not

even aware that the task constitutes a test. This approach has been used in developing an assessment scheme for one of the curriculum projects described by Harmon and Mokros. In that project, for example, children are asked to draw a circuit as part of the electricity unit. The resulting drawings are collected and forwarded to the curriculum developers as part of their assessment system.

Besides the continuum of assessment methods, and interrelated with it, is a continuum of assessment *uses*:

a) *Classroom management/pedagogic use*: One purpose of assessment is to help teachers operate their classrooms efficiently; that is, to provide appropriate activities for children, activities that make sure they progress and are profitably occupied. This is a particularly important task in any classroom in which the teacher is responsive to individual children's needs and involves the children with materials. Obviously, a teacher who teaches set lessons regardless of the students' responses does not have this need. At the level of the individual child, the most important kind of assessment methods are those included under record keeping. However, all teachers who organize and direct elementary classrooms in which materials-based science is used need to and do keep track of what the children are doing. More or less formally, they keep these records.

b) *Diagnosis, individual evaluation, and classroom-level evaluation*: At this level of assessment, documentation can play a key role. In order for teachers to understand and appreciate the growth that students show, in order for them to provide for individual children and to be able to discuss intelligently both the successes and the problems children face, teachers need viable documentary evidence. Besides the tests that are so commonly used, a wide range of methods can provide this information, especially when data are collected over some period of time. Documentation can also serve as the basis for setting policy. Insight into children's progress and into the factors that may cause them difficulty can be illuminating to staff at the classroom, school, or system level in modifying curriculum, providing needed services, or justifying a curriculum or set of activities.

c) Finally, for purposes of *large-scale* policy, such as state or national surveys of science achievements, testing, as it has been broadly defined above, is often the most general and most useful method.

The overlap of assessment methods and assessment purposes provides a parallel set of continua that is summarized in table 1.

TABLE 1. Assessment Methods and Purposes

	most particular	↔	most general
methods	record keeping	↔	documentation ↔ testing
purposes	classroom management	↔	student progress ↔ policy

Within this framework, we can make the following generalizations:

- We need to emphasize the versatility, power, and value of the range of documentation methods. We should publicize and elaborate documentable activities: kids' questions, hypotheses and metaphors, classroom discussions, and the products of science activity.
- We need to elaborate methods for summarizing and presenting the results of documentation such as charts and graphs that illustrate what individuals and groups of children have demonstrated. In the field of reading assessment, we are beginning to see the results of such documentation summaries, as evaluators present tables, charts, and graphic displays that summarize children's progress as demonstrated through the documentation efforts. A similar development can be anticipated in science assessment as the expanded use of a wider range of assessment methods becomes available.
- "Displays" of science achievement (usually the result of some documentable activity) can be powerful. When presented to parents, administrators, and school boards, they can persuade decision makers of the value of doing science. Again, we need to document this form of documentation.
- We need to expand the repertoire of "finding out" modes of assessment. Methods that tell us something about how learners grapple with problems, rather than what they know, need to be emphasized and exploited. The last four papers in this volume illustrate how much an expanded repertoire of assessment methods can teach us about how children learn.

- The analysis of types of assessment and their uses allows a discussion of the best way to apply the strengths of each method. If testing is primarily seen as a methodology for large-scale comparisons and system-wide “checking-up,” then, clearly, not every child needs to be tested on each question; matrix sampling and other methods that spread the testing burden become realistic. The conference attendees all recognized that assessment would improve if we stopped testing every child with the kinds of tests we use now.

ASSESSMENT AND INSTRUCTION

A topic that comes up repeatedly in the papers in this volume, and which was discussed at length during the conference, is the relationship between assessment and instruction; the match (or lack of it) between evaluation and pedagogy. Assessments that are limited primarily to short-answer and multiple-choice questions provide a particular view of what science is, a view at odds with most reasonable descriptions of this subject. This topic receives the most detailed treatment in Frank Davis’s chapter, in which he contrasts various philosophies of education and then suggests how each of them leads to different assessment strategies.

In other papers, the relationship is sometimes portrayed by discussing the mismatch between current assessment approaches and pedagogic goals. The curriculum developers describe this dichotomy clearly, with language that is reminiscent of the kind of disparities between their approaches to education and the structure of schools that open-education advocates discussed 20 years ago (DeRivera, 1970).

This disparity becomes most apparent when test questions deviate from simple recall of facts and attempt to probe more complex components of the scientific process. For example, one type of question asks the student to provide the best answer among a set of possible ones. All questions of this type skew the nature of scientific reasoning. In science, “best” methods, or even best explanations, don’t exist in the absence of a context.

Alternatively, questions of this type purport to assess higher-order thinking skills but are either ambiguous or test simple knowledge. For example, consider this question reprinted in the 1988 NAEP report (Mullis and Jenkins, 1988):

Which of the following best explains why marine algae are most often restricted to the top 100 meters of the ocean?

- ◊ They have no roots to anchor them to the ocean floor
- ◊ They are photosynthetic and can live only where there is light

- ◇ The pressure is too great for them to survive below 100 meters
- ◇ The temperature of the top 100 meters of ocean is ideal for them.

This question is intended to assess whether students can “analyze scientific procedures and data.” But each possible answer is amenable to experiment, so deciding which is the “best” answer cannot come from analysis of procedures (or of data, since insufficient data are given), but only from recall of knowledge.

An often-cited example of the limitations of multiple-choice questions is the potential physics question about how a student might use a barometer to measure the height of a building. The possible answers included dropping the barometer off the roof and counting the seconds until it struck the ground and offering to trade the barometer to the janitor in return for the desired information. Are these methods less “good”—less accurate, less moral, or less reliable—than methods that use the atmospheric pressure measuring ability of the barometer? Who is to say?

The use of such “best” answer questions suggests to me a parallel with the feminist critique of moral-development schemes. Kohlberg (1973) and associates argued that people may advance to increasingly higher moral stages as they mature, progressing from an egocentric stage through conditional stages to a stage where basic moral principles govern their actions. Carol Gilligan (1982) and others have argued that this view of morality arbitrarily determines that an abstract, impersonal application of moral laws is of “higher” value than a situational one that might consider specific circumstances—the feelings of individuals, for example. The feminist critics go on to point out that women are more likely than men to have been socialized to consider such factors and when confronted with moral dilemma problems (a kind of test) on the average provide answers different from those given by men. No generally accepted external criterion, except one imposed by a particular group of researchers (who made up the test) can be used to conclude that reasoning leading to one type of answer is of higher value than another type of reasoning.

Similarly, in the case of scientific reasoning, there is little in the way of external criteria for many of the problems that are presented in tests to determine whether one type of reasoning, experimental approach, or method of inquiry is “best.” That conclusion depends on the thinking of those who have developed the test. In any real situation, the best method is the most expedient for providing reliable results. Scientists sometimes speak of “elegant” solutions, or “elegant” procedures: these may not be better but are aesthetically more appealing. But “best” cannot be applied in the abstract to a single question taken out of a context of use. The

famous dictum, "Science is doing your damndest with your mind, no holds barred," certainly applies here. The use of questions that ask for the "best" answer makes an incorrect statement about the nature of science.

POLITICAL IMPLICATIONS OF TYPES OF ASSESSMENT

The mismatch between traditional assessment and the nature of science is not the only foundation upon which the argument concerning the need for more diversified assessment methods rests. Several of the authors advance arguments that I would generally classify as political. The style of assessment makes a statement about the fundamental political nature of the educational system we advocate.

Assessments that are separate from teaching, that reduce student involvement to passive response to questions framed outside the local world of the classroom and its activities, and that value only some kinds of in-school learning, suggest strongly that the process of education is authoritarian, dehumanizing, and not empowering to the children who partake of it. Engel points out that traditional assessment methods lead to feelings of failure, substitute external rewards for intrinsic motivation, and set up systems of competition that replace the rewards of mastery.

Alternative assessment methods are also required to meet the professional needs of teachers and to use the unique knowledge of children that teachers amass during their interaction with pupils. One consequence of increased reliance on national, impersonal tests is the devaluation of teachers' skills as professionals. As these assessment methods increasingly dominate the schools, teachers are less able to put their individual stamp on the pedagogic process, and their knowledge of what children can and cannot do becomes increasingly irrelevant to the educational process. Patricia Stock, Maryellen Harmon and Jan Mokros, Hubert Dyasi, and Edward Chittenden all stress the role that the individual teacher can and should play in assessment. This role is enhanced as record-keeping and documentation processes play an increasing part in assessment.

Finally, the assessment approaches we choose reflect the value we place on each child. The more standardized the test, the more likely that it will not adequately serve individual children. This point is so obvious that perhaps it should be unnecessary to make. Any assessment based on a small number of items that do not vary in content, style, or method of testing and that require only a depersonalized response (filling in one of four circles) enormously increases the probability that any individual child's result will not reflect that child's abilities. There are, in the United

States, children who have difficulty with English; who have physical handicaps; who think more complexly than is envisioned by any particular test maker; who have personal associations with particular pictures, diagrams, or graphs that alter their responses; or who, for a range of cultural and ethnic reasons, react differently from the norm to any single assessment methodology. Each of these circumstances has been documented in a variety of publications critical of multiple-choice tests.

By limiting our assessment methods to one form, we will certainly mis-assess some children. In those situations where only national trends or regional achievements are at issue, this lack of proper representation for each child may be insignificant. That is why tests are sometimes appropriate for national surveys and why not every child needs to be assessed. But when these tests are then also used for diagnostic purposes, for placement of children, or to determine grades, we can be sure that significant numbers of children will be penalized, and others unfairly rewarded.

I have called many of the issues involved in matching assessment and instruction political issues because the methods of assessment we use have political implications. What we do makes a value statement about our view of teachers and students and about the very nature of education. But these same issues could also be called pedagogic ones—they reflect our professional judgment about the nature of the teaching and learning processes. Are students passive or active learners? Are teachers technicians who pass on canonical views of science or active participants in the construction of knowledge in the classroom? Is each child unique, with a separate learning style and a background of experience he or she brings to the classroom, or are children standard vessels to be filled with knowledge? The range of assessment methods we allow in our classrooms and our schools says more about our answers to these questions than any statement of faith we may utter to parents, print in school department reports, or assert in papers for inservice courses.

DEVELOPMENTAL ISSUES: STUDENT PROGRESS

The reliance in the past on short-answer, multiple-choice assessment measures has obscured an important component of science development: children's progress along a continuum of increasing knowledge of science processes and skills. Brenda Engel points out that the need to convert learning into a score and a grade based on tests can actually diminish the progress a child makes. Similarly, the emphasis on those components of science that are testable by that style of assessment has caused the science education community to devote insufficient resources to elements

of science education that are related to development. Driver points out that we need to understand how children make progress (indeed, we need to define what progress is in conceptual development), then we need to develop science lessons that contribute to that progress and assess accordingly. Any assessment scheme must take into account not only the ultimate outcome of changes in children's world views, but also the intellectual movement toward change. A student who believes that heavy objects fall faster than light ones is "wrong," but may express this incorrect notion in various ways. One student may marshal evidence in support of this view (a stone does fall faster than a feather in air); another may struggle with this view but not express it clearly; another may simply state it for reasons she cannot explain. We need to develop systems that can distinguish between various stages of understanding.

Systems of documentation that represent the middle ground of assessment can contribute to this understanding of children's understanding. This is especially true if we use these systems over long time periods. The kinds of arguments that Sally gives in the second grade may be different from the kinds of arguments she puts forward in the third grade. The kinds of experiments she attempts may vary from fourth to sixth grades. One of the characteristics of assessment systems that involve students in performing experiments and expressing their views is that the assessors begin to be able to put together a picture of what individual children and groups of children are and are not able to accomplish at various stages. The APU experience provides evidence of what typical 11-year-olds in England are capable of doing, how their scientific beliefs intersect with their scientific actions.

In the last decade, we have learned a great deal about children's understanding of science concepts. The work Driver summarizes illustrates just how much. Yet, we know comparatively little about how children's science skills develop. How does one become a good observer, a good measurer, or a good formulator of questions? Even more important, what do we mean when we describe someone as possessing these skills? Especially, what do we mean by these terms when applied to children of different ages? Surely a careful observation of an event in nature by a kindergarten child is not the same as what we could expect from a sixth grader.

But there has been too little research on the development of science skills in elementary school children. More significant for this discussion, the development of these skills is hardly recognized in the descriptions of science provided by states and local

school districts. For example, the New York State Curriculum (New York, 1987) outlines various levels of content knowledge (increasingly complex content is proposed as children progress from level I, ages 4-7 years, to level III, ages 9-11) but describes process skills only once, with no age differentiation.

Similarly, the Boston Public Schools' elementary science curriculum (Marshall, 1983), which describes specific activities for each skill and content area, defines observation skills almost identically for first and sixth grades, as follows:

- **FIRST GRADE; OBSERVING:** Observing carefully; noticing such details as size, shape, color, texture, weight, taste, smell, position, time taken, etc. (possibly using a hand lens.)
- **SIXTH GRADE; OBSERVING:** Observing carefully; noticing such details as size, shape, color, texture, weight, smell, taste, position, time taken, etc. (using a magnifying glass or microscope when necessary.)

Two separate but related developmental concerns are expressed in the papers and in the discussions.

First is the need for assessment over time. Pine argues that any independent measure to calibrate science assessments must be based on the judgment of someone (most likely the child's teacher) who has observed and documented work over a long time. Both Engel and Stock, in their descriptions of assessment in the language arts, stress the need to understand a student's work as it evolves over a period of time. The methods Dyasi describes require collecting data over time. Some aspects of assessment simply cannot be carried out with only a single snapshot of a child's performance.

The second aspect of time involved concerns the instructional activities that occur during the time included in long-term assessments. We conjecture that a more accurate picture of a child's progress can be achieved by stretching out the assessment period to encompass months or years, but this will be true only if significant time is spent in science activities during this period. It does no good to compare a child's performance from one year to the next in, for example, playing a musical instrument, unless the child has practiced and received instruction during the interim. Science education in most elementary schools receives so little time, usually less than an hour a week, that it is unreasonable to expect much improvement in any serious assessment. It is a wonder that our children appear to progress at all. Given the dearth of instruction, the apparent progress children make on assessment

tests raises questions about the tests themselves (Buros, 1977). Some teachers devote considerable time and effort to science, and their efforts are usually apparent to any interested observer. When children have the opportunity to do science, they become familiar with both the processes and the content, and they display their knowledge in a variety of ways. The extraordinary drawings that accompany Chittenden's paper are an example. There can be no doubt that these first-grade children spent many hours observing monarch butterflies in various stages of their life cycle. Some of the examples Dyasi provides also illustrate the children's familiarity with the objects they classify. Nothing can substitute for time spent on science; unless we provide that time for our children, no comprehensive assessment plan will do more than document the inadequacy of their education.

CONCLUSION

The dominant mode of assessment for science education in the United States today is based on inadequate paper-and-pencil, short-answer (usually multiple-choice) tests. If we are to improve science education, we must begin to move away from this model, just as literacy and writing assessments have expanded to include a wider set of methods.

There is considerable evidence that movement in this direction has begun. Every major study in recent years has stressed the need for research related to assessment and development of more comprehensive assessment methods:

- The planning study commissioned by the National Science Foundation (Knapp, et al., 1987) lists improvements in mathematics and science assessment as one of the major areas for NSF investment.
- The National Research Council (NRC) Committee on Research in Mathematics, Science and Technology Education, in its publication *Education and Learning to Think* (Resnick, 1987), argues that a new generation of assessment instruments is needed in order to assess higher-order thinking skills adequately.
- Many studies have demonstrated that the currently available tests used by schools are inadequate and inappropriate for the assessment of meaningful science learning (see Hein, 1987, for a recent discussion).
- The first recommendation in the second report (Murnane and Raizen, 1988) of the Committee on Indicators of Precollege Science and Mathematics Education of the NRC is:

Research and Development: To provide the requisite tests for use as indicators of student learning, the committee recommends that a greatly accelerated program of research and development be undertaken aimed at the construction of free-response materials and techniques that measure skills not measured with multiple choice tests. The committee urges that the development of science tests at the K - 5 level receive immediate attention (p. 5).

- In *Assessment in Elementary School Science Education* (Raizen, et al., 1989), the authors, in recommending more research in support of assessment, argue that “by increasing their understanding of how children learn science, educators could dramatically improve instructional effectiveness....”

Similar views are expressed by Jones (1989) in another review of science and mathematics achievement.

In addition, a number of state assessment programs have begun to institute tests that go beyond traditional measures. New York State and Massachusetts have recently added performance measures to their statewide science testing, California is planning to institute both performance tests and written questions that require narrative answers, and Connecticut has proposed, in cooperation with several other states, a dramatically different form of science assessment, including portfolios and other collections of materials, as well as performance tests.

Finally, as indicated in this volume, researchers and curriculum developers are actively engaged in broadening our knowledge of assessment methods and in incorporating alternative forms of assessment into curriculum materials.

If we want to change the way we teach science, we have to simultaneously change the way we assess science. This task is difficult and complex; it requires the cooperative efforts of many professionals. Any single professional group always brings to bear on a problem both the strengths and the limits of its world view. Science assessment is too important to be left in the hands of only psychometricians and other test developers. As a problem that combines theoretical and real-world issues, it requires input from groups representing many perspectives. Only then will we come to solutions that have both theoretical validity and application in the real world of schools.

Notes

PART ONE: LESSONS FROM THE ASSESSMENT OF READING AND WRITING

1. Lessons from Literacy

1. For an explanation of the distinction between acquisition of knowledge and learning—more particularly, between acquisition of language which everyone experiences and learning about language, which linguists do—see Gee, James Paul, “What is Literacy?” *Teaching and Learning, the Journal of Natural Inquiry*, Vol. 2, No. 1, Fall 1987.
2. Along with predictability, generalizability, and mathematization, objectivity is one of the largely unexamined assumptions informing most evaluation methods.
3. This kind of failure is partly responsible for the currently perceived need to teach “thinking skills” in the upper elementary grades.

2. Taking on Testing: Chapter Two

1. These arguments may also be found in a series of unpublished papers written by teachers of English in the public schools of the city of Ann Arbor, Michigan, and available through the Center for Educational Improvement through Collaboration (CEIC), 2014 SEB, The University of Michigan, Ann Arbor, MI 48109. Specifically, see: C. Susan Frazier, *The Student as Writer*; Jean Long, *The Context and the Occasion*; and David Stringer, *The Teacher as Reader*.
2. The following teacher-researchers participated in our project to develop a local assessment of writing in Saginaw: From the schools—Linda Bush, Jean Cole, Alena Dancy, Jane Denton, Sharon Floyd, Louise Harrison, Robert Hoard, James Jones, Mary Lane, Gail Oliver, Mary Roberts, Kathie Smith, Sheila Smith, Bea Ugartechea, Rosa Winchester, and Carol Woolfolk; from the University of Michigan—Cathy Fleischer, Richard Harmston, Jay Robinson, David Schaafsma, and Patricia Stock.

Those of us who worked together particularly wish to thank and commend others in the school district whose constant support

of our work made it possible: Foster B. Gibbs, Superintendent; Burris Smith, Director, K-12 Education; Gene Nuckolls, Assistant Superintendent, Secondary Education; Lochie Overbey, Coordinator, Staff Development; James Jones, former Language Arts Coordinator; Jane Denton, Language Arts Coordinator; Thomas Sharp, Principal, Arthur Hill High School; Wilson Smith, Principal, Saginaw High School.

PART TWO: ASSESSMENT THEORY

3. Assessment and Teaching of Thinking Skills

1. Recently, while presenting a seminar titled "A taxonomy of higher-order thinking skills," I pointed out to the audience that my first transparency used the term higher-order cognitive skills. I commented that the difference is not significant. The expressions on the faces caused me to ask, "Or are they?" An extended and vigorous discussion ensued among the assembled educational and psychological researchers delineating differences in meaning that individuals attributed to thinking and cognitive skills. In another situation in which I was using reasoning skills to describe an outcome of formal education to a group of academicians, several members of the group expressed discomfort with the phrase and insisted on substituting intellectual skills for reasoning skills.
2. An example of the way in which performance and thinking skills are undifferentiated is illustrated in an article that recently crossed my desk describing a resource available for teachers to incorporate thinking skills into regular school curricula. The thinking skills the materials were designed to teach included: the accuracy and detail with which students can describe an interruption of a class by a stranger, creative problem solving, problem solving skills, control of variables, decision making, and critical thinking.
3. The hands-on version of this task is one of the exercises designed and field tested as part of a project conducted by the ETS to study methods for assessing higher order thinking skills in science and mathematics.

4. An example of an item to assess factual information:

The planet closest to the sun is:

Earth
Mercury
Pluto
Uranus
Venus

5. Assessing Science Education: A Case for Multiple Perspectives

1. References for many of the works cited in this paper can be found in my dissertation, *How Do Adults Learn?: Metatheory of Learning Based on Stephen Pepper's Theory of World Hypotheses*, available through University Microfilms, Ann Arbor, Michigan.

PART THREE: LARGE-SCALE ASSESSMENTS

6. What We Learn from State Assessments of Elementary School Science

1. Many of these ideas have previously appeared in the Connecticut Assessment of Educational Progress 1984-85 Science Summary and Interpretations Report, State of Connecticut, department of Education, Hartford, Connecticut, 1986.
2. The Assessment Performance Unit (APU) in Great Britain has served as a model and inspiration for Connecticut's work in performance assessment. We are grateful to the APU for letting us begin with their exercises and modify them as necessary for use in Connecticut.
3. Connecticut uses holistic scoring to score 100,000 papers for students in grades 4, 6, and 8. Certified Connecticut teachers are trained at a central site and work for five-and-one-half hours per day. The rater gives each paper an overall rating, and, to ensure interrater reliability, each paper is read independently by two raters. Raters average approximately 30 papers an hour (including training time), which costs less than \$1.00 per student for two scorers to independently determine a score for a student's paper and encode the score onto a preprinted machine-scorable answer sheet.
4. The total cost of administering the practical testing in Connecticut was \$6,000. The sample included 300 students at grades 4, 8, and 11 for a total of 900 students. The cost per student was \$6.66. This cost includes developing the scoring rubrics and training external administrators to administer the performance items and encode the students' responses on a preprinted machine-scorable answer sheet. It does not include the cost of developing the tasks or printing the machine-scorable answer sheets.

7. What Has Been Learnt About Assessment from the Work of the APU Science Project?

1. Acknowledgement is due to the Department of Education and Science for permission to include the data in this paper. Any views expressed are those of the author.

PART FOUR: ASSESSMENT IN SCIENCE EDUCATION RESEARCH AND DEVELOPMENT

8. Assessment in the New NSF Elementary Science Curricula: An Emerging Role

1. Both the 1983 and 1988 NAEP reports indicate that no more than 50 percent of the teachers in grade 3 teach science for 1-2 hours per week, and 21 percent do so for far less. In 1977, Weiss reported that less than 18 minutes per day was spent on science in K-3 as compared with 40 minutes in mathematics and 75 minutes in reading; in grades 4-6, the average was 29 minutes per day for science. Her 1986 study revealed little change. But it must be noted that these data are drawn from teachers' self-reporting. Mullis and Jenkins present a less optimistic picture.

2. *Embedded assessment* is the term we have chosen to use for learning experiences that serve a double function: a discovery experience for students but written in such a way as to free the teacher to observe carefully student scientific processes and application of concepts. An *extended problem* is a real-world problem-solving task that is undertaken by groups of students and extends over several days.

3. By *justified multiple choice* we mean multiple-choice questions that require the student to justify in one or two lines why this choice is the correct one or why the other alternatives are less good.

Bibliography

INTRODUCTION

Assessing Assessment

- Duckworth, E. 1978. *The African Primary Science Program: An Evaluation*. Grand Forks, ND: North Dakota Study Group on Evaluation, University of North Dakota Press.
- Hawkins, D. 1974. On Living in Trees. In *The Informed Vision*. New York: Agathon Press.
- Hoepfner, R., et al. 1976. *CSE Elementary School Test Evaluations*. Los Angeles: Center for the Study of Evaluation, U.C.L.A.
- Miller, J.D. 1988. The Roots of Scientific Literacy: The Role of Informal Learning. In *Science Learning in the Informal Setting*, ed. P.G. Heltne and L.A. Marquard. Chicago: The Chicago Academy of Sciences.
- Mullis, I.V.S. and L.B. Jenkins. 1988. *The Science Report Card-Elements of Risk and Recovery*. Princeton, NJ: Educational Testing Service.
- National Assessment of Educational Progress. 1987. Report No:17-HOS-80. *Learning by Doing*. Princeton, NJ: Educational Testing Service.
- Novak, J. 1987. Proceedings of the Second International Seminar, Misconceptions and Educational Strategies in Science and Mathematics. Ithaca, NY: Cornell University.
- Resnick, L.B. 1987. Learning in school and out. *Educational Researcher* 16(9): 13-20.

PART ONE: LESSONS FROM THE ASSESSMENT OF READING AND WRITING

1. Lessons from Literacy

- Clay, M.M. 1985. *The Early Detection of Reading Difficulties*. Portsmouth, NH: Heinemann.
- Clay, M.M. 1972. *The Early Detection of Reading Difficulties, a Diagnostic Survey with Recovery Procedures*. New Zealand: Heinemann Education.

- Ferreiro, E. and A. Teberosky. 1982. *Literacy Before Schooling*. Portsmouth, NH: Heinemann.
- Gee, J.P. 1987 (Fall). What is literacy? *Teaching and Learning, The Journal of Natural Inquiry* 2(1).
- Goodman, K., ed. 1973. *Miscue Analysis, Applications to Reading Instruction*. Detroit: Wayne State University.
- Goodman, Y.M. and B. Altwerger. 1981. *Print Awareness in Pre-School Children: A Working Paper*. Arizona: Arizona Center for Research and Development.
- Harste, J.C., V. Woodward and C.L. Burke. 1984. *Language Stories and Literacy Lessons*. Portsmouth, NH: Heinemann.
- Holdaway, D. 1979. *The Foundations of Literacy*. Sydney: Ashton Scholastic.
- National Association of State School Boards. 1988. *Right From the Start*.

2. Taking on Testing: Chapter Two

- Barritt, L., et al. 1986. Researching Practice: Evaluating Assessment Essays. *College Composition and Communication* XXXVII (3): 315-27.
- Clark, M. 1983. Evaluating Writing in an Academic Setting. In *Forum: Essays on Theory and Practice in the Teaching of Writing*, ed. P.L. Stock, 59-79. Portsmouth, NH: Heinemann Boynton/Cook.
- Cooper, C.R., ed. 1981. *The Nature and Measurement of Competency in English*. Urbana, IL: NCTE.
- Cooper, C.R. and C.L. Odell, eds. 1977. *Evaluating Writing: Describing, Measuring, Judging*. Urbana, IL: NCTE.
- Heath, S.B. and A. Branscombe. 1985. Intelligent Writing in an Audience Community: Teacher, Students, and Researchers. In *The Acquisition of Written Language*, ed. S.W. Freedman, 3-32. Norwood, NJ: Ablex.
- Robinson, J.L. Literacy and Conversation: Notes Toward a Constitutive Rhetoric. Unpublished paper.
- Stock, P.L. and J.L. Robinson. 1987. Taking on Testing: Teachers as Tester-Researchers. *English Education* 19(2): 93-121.
- White, E.M. 1985. *Teaching and Assessing Writing*. San Francisco: Jossey-Bass.
- Wixon, V. and P. Wixon. 1977. Getting it Out, Getting it Down: Adapting Zoellner's Talk-Write. *English Journal* 66: 70-73.

PART TWO: ASSESSMENT THEORY

Introduction

Wiggins, G. 1989. A True Test: Towards More Authentic and Equitable Assessment. *Phi Delta Kappan* 70: 703-713.

3. Assessment and Teaching of Thinking Skills

- Anamuah-Mensah, J. 1986. Cognitive strategies used by chemistry students to solve volumetric analysis problems. *Journal of Research in Science Teaching* 23 (December): 759-69.
- Atwater, M.M. and R.D. Simpson. 1984. Cognitive and affective variables affecting black freshmen in science and engineering at a predominantly white university. *School Science and Mathematics* 84 (February): 100-112.
- Berger, C.F. and P.R. Pintrich. 1986. Attainment of skill in using science processes: grade and task effects. *Journal of Research in Science Teaching* 23 (November): 739-47.
- Bodner, G.M. and T.L.B. McMillen. 1986. Cognitive restructuring as an early stage in problem solving. *Journal of Research in Science Teaching* 23 (November): 727-37.
- Boreham, N.C. 1985. The effect of sequence of instruction on students' cognitive preferences and recall in the context of a problem-oriented method of teaching. *Instructional Science* 13 (March): 329-45.
- Burns, J.C., J.R. Okey and K.C. Wise. 1985. Development of Integrated Process Skills Test: TIPS. *Journal of Research in Science Teaching* 22 (February): 169-77.
- Champagne, A.B. (Undated). *Research Matters...to the Science Teacher: Definition and Assessment of the Higher Order Cognitive Skills*. National Association for Research in Science Teaching. Washington, DC: National Science Teachers Association.
- Champagne, A.B. and J. Rogalska-Saz. 1984 (April). *Learning to borrow: A theoretical analysis of traditional subtraction instruction*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.
- Chandran, S.H., D.F. Treagust and K. Tobin. 1987. The role of cognitive factors in chemistry achievement. *Journal of Research in Science Teaching* 24 (February): 145-60.
- Chittenden, E. and D. Meier. 1985. Science testing: creative alternatives. *Curriculum* 25 (September/October): 76-77.
- Collis, K.F. and H.A. Davey. 1986. A technique for evaluating skills in high school science. *Journal of Research in Science Teaching* 23 (October): 651-63.

- Conwell, C.R., S.L. Helgelson, and D.G. Wachowiak. 1987. The effect of matching and mismatching cognitive style and science education. *Journal of Research in Science Teaching* 24: 713-22.
- De Jong, T. and M.G.M. Ferguson-Hessle. 1986. Cognitive structures of good and poor novice problem solvers in physics. *Educational Psychologist* 78 (August): 279-88.
- Embretson, S. 1988 (August). Cognitive Design Systems for Psychological Tests. Invited address. Mathematical and Statistical Models of Behavior session presented at annual meeting of American Psychological Association, Atlanta, GA.
- Fourier, M.J. 1983. Academic achievement of students who receive disclosure of cognitive style map information. *Journal of Experimental Education* 51 (Spring): 122-30.
- Harty, H. and D. Beall. 1984. Toward the development of a children's science curiosity measure. *Journal of Research in Science Teaching* 21: 425-36.
- Hassan, A.M.A. and R.L. Shrigley. 1984. Designing a Likert Scale to measure chemistry attitudes. *School Science and Mathematics* 84 (December): 659-69.
- Heath, P.A., A.L. White, D.F. Berlin and J.C. Park. 1987. Decision making: influence of features and presentation mode upon generation of alternatives. *Journal of Research in Science Teaching* 24 (December): 821-33.
- Hein, G.E. 1987. The right test for hands-on learning. *Science and Children* 25(2): 8-12.
- Heller, P.M., et al. (in press). Proportional reasoning: the effect of two context variables, rate type and problem setting. *Journal of Research in Science Teaching*.
- Kavale, K.A. and S.R. Forness. 1987. Substance overstyle: assessing the efficacy of modality testing and teaching. *Exceptional Children* 54 (November): 228-39.
- O'Donnell, A.M. 1987. Cognitive social/affective and metacognitive outcomes of scripted cooperative learning. *Journal of Educational Psychology* 79 (December): 431-437.
- Romberg, T.A. and K.F. Collis. 1985. Cognitive functioning and performance on addition and subtraction word problems. *Journal of Research in Mathematics Education* 16 (November): 375-82.
- Saunders, W.L. and J. Jesunathadas. 1988. The effect of task content upon proportional reasoning. *Journal of Research in Science Teaching* 25 (January): 59-67.

- Singh, B. 1986-87. The development of tests to measure mathematical creativity. *International Journal of Mathematical Education in Science and Technology* 18: 181-6.
- Thompson, C.L. and R.H. Shrigley. 1986. What research says: revising the science attitude scale. *School Science and Mathematics* 86 (April): 331-43.
- Trifone, J.D. 1987. The test of logical thinking. *The American Biology Teacher* 49 (November/December): 411-416.
- Wittig, M.A., S.H. Sasse and J. Giacomini. 1984. Predictive validity of five cognitive skills tests among women receiving engineering training. *Journal of Research in Science Teaching* 21 (May): 537-46.
- Yager, R.E. 1987. Assess all five domains of science. *Science Teacher* 54 (October): 33-7.

4. Validity of Science Assessments

- Assessment of Performance Unit. 1985. *Science in Schools, Age 11 Report No. 4*. London: Department of Education and Science.
- California Assessment Program. 1986. *1985-86 Annual Report*. Sacramento: California State Dept. of Education.
- College Entrance Examination Board. 1988. *10 SAT's*. 3d ed. Princeton: College Entrance Examination Board.
- Messick, S. 1989. Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher* 8: 5-11.
- Murnane, R. and S. Raizen, eds. 1988. *Improving Indicators of the Quality of Science and Mathematics Education in Grades K-12*. Washington, DC: National Academy Press.
- Task Group on Assessment and Testing. 1988. *Report to the Secretary of State for Education and Science*. London: Department of Education and Science and the Welsh Office, Her Majesty's Stationery Office.
- Wall, J. 1981. *Compendium of Standardized Science Tests*. Washington, DC: National Science Teachers Association.

5. Assessing Science Education: A Case for Multiple Perspectives

- Davis, F. *How Do Adults Learn?: Metatheory of Learning Based on Stephen Pepper's Theory of World Hypotheses*. Ann Arbor, MI: University Microfilms.
- Pepper, S. 1942. *World Hypotheses: A Study in Evidence*. Berkeley: University of California Press.

PART THREE: LARGE-SCALE ASSESSMENTS

6. What We Learn from State Assessments of Elementary School Science

- American Association for the Advancement of Science. 1989. *Project 2061: Science for all Americans*. Washington, DC: AAAS.
- Baron, J.B., et al. 1989. Toward a new generation of student outcome measures: Connecticut's Common Core of Learning Assessment. Presented at the American Educational Research Association Annual Meeting, San Francisco, CA.
- Connecticut State Department of Education. 1986. *Connecticut assessment of educational progress 1984-85: science summary and interpretations*. Hartford, CT: State of Connecticut, Department of Education.
- International Association for the Evaluation of Educational Achievement. 1988. *Achievement in Seventeen Countries*. Oxford, England: Pergamon Press.
- National Assessment of Educational Progress. 1987. *Learning by doing: a manual for teaching and assessing higher-order thinking in science and mathematics*. Princeton, NJ: Educational Testing Service.
- Pecheone, R.L., et al. 1988. A Comprehensive Approach to Teacher Assessment: Examples From Math and Science. In *Science Teaching: Making the System Work*, ed. A.B. Champagne. Washington, DC: American Association for the Advancement of Science.
- Raizen, S., et al. 1989. *Assessment in Elementary School Science Education*. Washington, DC: The National Center for Improving Science Education (Publication No. 303).

7. What Has Been Learnt about Assessment from the Work of the APU Science Project?

- DES. 1974. *Educational Disadvantage and the Educational Needs of Immigrants*. (Cmnd 5720) London: HMSO.
- DES. 1978. *Science Progress Report 1977-8*. London: DES.
- DES. 1981a. *Science in Schools. Age 11: Report No. 1*. London: HMSO.
- DES. 1982b. *Science in Schools. Age 13: Report No. 1*. London: HMSO.
- DES. 1982c. *Science in Schools. Age 15: Report No. 1*. London: HMSO.

- DES. 1983a. *Science in Schools. Age 11: Report No. 2*. London: DES.
- DES. 1984a. *Science in Schools. Age 11: Report No. 3*. London: DES.
- DES. 1984b. *Science in Schools. Age 13: Report No. 2*. London: DES.
- DES. 1984c. *Science in Schools. Age 15: Report No. 2*. London: DES.
- DES. 1985. *Science 5-16: A Statement of Policy*. London: HMSO.
- DES. 1985a. *Science in Schools. Age 11: Report No. 4*. London: DES.
- DES. 1985b. *Science in Schools. Ages 13 and 15: Report No. 3*. London: DES.
- DES. 1986b. *Science in Schools. Age 13: Report No. 4*. London: DES.
- DES. 1986c. *Science in Schools. Age 15: Report No. 4*. London: DES.
- DES. 1987. *The National Curriculum 5-16.5. Consultation Document*. London: DES.
- DES. 1987. *Task Group on Assessment and Testing: A Report*. London: DES.
- DES. 1988a. *Science in Schools. Age 11: Review Report*. London: DES.
- DES. 1988b. *Science in Schools. Age 15: Review Report*. London: DES.
- DES. 1988c. *Science in Schools. Age 11, 13 and 15: A Technical Review Report*. London: DES.
- DES. 1989a. *Science in Schools. Age 13: Review Report*. London: DES.
- Donnolly, J. 1988. *Metals at Age 15. Science Report for Teachers: 10*. ASE
- Gamble, R., et al. 1985. *Science at age 15. Science Report for Teachers: 5*. ASE.
- Gott, R. and P. Murphy. 1987. *Assessing investigations in science, ages 13 and 15. Science Report for Teachers: 9*. ASE.
- Harlen, W. 1983. *Science at age 11. Science Report for Teachers: 11*. ASE
- Harlen, W. 1987. *Planning scientific investigations at age 11. Science Report for Teachers: 8*. ASE.

- Harlen, W., D. Palacio, and T. Russell. 1984. *The APU assessment framework for science at age 11. Science Report for Teachers: 4. ASE.*
- Johnson, S. and J.F. Bell. 1985. Evaluating and predicting survey efficiency using generalizability theory. *Journal of Educational Measurement* 22: 107-119.
- Murphy, P. 1987. Investigations in Science. Paper presented at the AERA Symposium, Children's Procedural Knowledge in Science. Washington, DC.
- Murphy, P. 1990. Gender and Assessment. In *Developments in Learning and Assessment*, eds. P. Murphy and B. Moon. Milton Keynes, England: Open University Press.
- Murphy, P. and R. Gott. 1984. *The assessment framework at ages 13 and 15. Science Report for Teachers: 2. ASE.*
- Murphy, P. and B. Schofield. 1984. *Science at age 13. Science Report for Teachers: 3. ASE.*
- National Curriculum Council. 1989. *Primary Science 1.*
- National Curriculum Council. 1989. *Primary Science 2.*
- National Curriculum Council. 1989. *Secondary Science.*
- Welford, G., W. Harlen, and B. Schofield. 1985. *Practical testing at ages 11, 13 and 15. Science Report for Teachers: 6. ASE.*

PART FOUR: ASSESSMENT IN SCIENCE EDUCATION RESEARCH AND DEVELOPMENT

8. Assessment in the New NSF Elementary Science Curricula: An Emerging Role

- Lamb, W., et al. 1989. *Physical Science*. Chicago: Harcourt Brace Jovanovich.
- Madaus, G.F. and D. Stufflebeam (eds). 1989. *Educational Evaluation: Classic Works of Ralph W. Tyler*. Boston: Kluwer Academic Publishers.
- Mullis, I.V.S. and L.B. Jenkins. 1988. *The Science Report Card: Elements of Risk and Recovery*. National Assessment for Educational Progress Report No 17-S-01. Princeton, NJ: Educational Testing Service.
- Raizen, S. 1988. *Assessment in Elementary Science Education*. In draft.
- Schwartz, J. 1977. Math Tests. In *The Myth of Measurability*, ed. P.L. Houts. New York: Hart Publishing Co.

- Taylor, E.F. 1977. Science Tests. In *The Myth of Measurability*, ed. P.L. Houts. New York: Hart Publishing Co.
- Weiss, I. 1977. *Report of the 1977 National Survey of Science, Mathematics, and Social Studies Education*. Washington DC: U.S. Government Printing Office SE 78-72.
- Weiss, I.S. 1987. *Report of the 1985-86 National Survey of Science and Mathematics Education*. Prepared for the National Science Foundation. No: SPE-8317070. Washington, DC: U.S. Government Printing Office.

9. Assessing the Progress of Children's Understanding in Science: A Developmental Perspective

- Bell, B. and A. Brook. 1985. *Aspects of secondary students' understanding of plant nutrition*. Children's Learning in Science Project, Centre for Studies in Science and Mathematics Education, University of Leeds.
- Bereiter, C. 1985. Toward a solution of the learning paradox. *Review of Educational Research* 55: 201-226.
- Brook, A. 1987. Designing Experiences to Take Account of the Development of Children's Ideas: An Example from the Teaching and Learning of Energy. In *Misconceptions and Educational Strategies in Science and Mathematics*. Proceedings of the Second International Seminar, ed. J. Novak. Ithaca, NY: Cornell University.
- Brook, A., H. Briggs and R. Driver. 1984. *Aspects of secondary students' understanding of the particulate nature of matter*. Children's Learning in Science Project, Centre for Studies in Science and Mathematics Education, University of Leeds.
- Brook, A. and R. Driver. 1988. *Progression in science: the development of students' understanding of physical characteristics of air, across the age range 5-16 years*. Children's Learning in Science Project Report. C.S.S.M.E., University of Leeds.
- Carey, S. 1986. Cognitive science and science education. *American Psychologist* 41(10): 1123-1130.
- Carey, S. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.
- Champagne, A.B., L.E. Klopfer and R.G. Gunstone. 1982. Cognitive research and the design of science instruction. *Educational Psychologist* 17(1): 31-52.
- Claxton, G. 1984. *Live and Learn*. New York: Harper and Row.

- Clement, J. and D. Brown. 1983. Using analogical reasoning to deal with 'deep' misconceptions in physics. Working paper, University of Massachusetts, Amherst.
- DiSessa, A. 1982. Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science* 6: 37-75.
- Donaldson, M. 1978. *Children's Minds*. London: Fontana/Collins.
- Driver, R. 1989. Changing conceptions. In: *Adolescent Development and School Science*, ed. P. Adey. Philadelphia: Falmer Press.
- Driver, R. and G. Erickson. 1983. Theories in action: Some theoretical and empirical issues in the study of students' conceptual frameworks in science. *Studies in Science Education* 10: 37-60.
- Driver R., E. Guesne and A. Tiberghien. 1985. *Children's Ideas in Science*. London: Open University Press.
- Driver R. and V. Oldham. 1986. A constructivist approach to curriculum development in science. *Studies in Science Education* 13: 105-122.
- Edwards, D. and N. Mercer. 1987. *Common Knowledge*. New York: Methuen.
- Engel-Clough, E. and R. Driver. 1986. Consistency in the use of students' conceptual frameworks across different task contexts. *Science Education*. 70(4): 473-496.
- Engel-Clough, E. and C. Wood-Robinson. 1986. Children's understanding of inheritance. *Journal of Biological Education* 19(4): 304-310.
- Erickson, G.L. 1979. Children's conceptions of heat and temperature. *Science Education* 63(2): 221-230.
- Erickson, G.L. 1980. Children's viewpoints of heat: a second look. *Science Education* 64: 323-336.
- Gilbert, J.K. and D.M. Watts. 1983. Concepts, misconceptions and alternative conceptions: changing perspectives in science education. *Studies in Science Education* 10: 61-98.
- Guesne, E. 1984. Children's ideas about light. *New Trends in Physics Teaching* IV. UNESCO.
- Gunstone, R. and R. White. 1981. Understanding of gravity. *Science Education* 65(3): 291-299.
- Helm, H. and J. Novak. 1983. *Proceedings of the International Seminar: Misconceptions in Science and Mathematics*. Ithaca: Cornell University.
- Hewson, P. and H. Hewson. 1984. The role of conceptual conflict in conceptual change and the design of science instruction. *Instructional Science* 13.

- Holding, B. 1987. Investigation of schoolchildren's understanding of the process of dissolving with special reference to the conservation of matter and the development of atomistic ideas. Ph.D diss., University of Leeds.
- Jung, W., H. Pfundt and C. Rhoneck, eds. 1982. *Problems concerning students' representation of physics and chemistry knowledge*. Proceedings of the international workshop. Ludwigsburg.
- Lawson, A.E. 1985. A review of research on formal reasoning and science teaching. *Journal of Research on Science Teaching* 22 (7)d: 569-617.
- McCloskey, M. 1983. Intuitive physics. *Scientific American* 248: 122-130.
- Minstrell, J. 1982. Explaining the 'at rest' condition of an object. *Physics Teacher* 203: 10.
- Nussbaum, J. 1985. The Earth as a Cosmic Body. In *Children's Ideas in Science*, ed. R. Driver, E. Guesne, and A. Tiberghien. London: Open University Press.
- Nussbaum, J. and S. Novick. 1982. Alternative frameworks, conceptual conflict and accommodation: Toward a principled teaching strategy. *Instructional Science* 11: 183-200.
- Osborne, R. and P. Freyberg. 1985. *Learning in Science*. Portsmouth, NH: Heinemann.
- Pfundt and Duit. 1985. *Bibliography: Students' Alternative Frameworks and Science Education*. Kiel: IPN.
- Posner, G.J., K.A. Strike, P.W. Hewson and Gertzog W.A. 1982. Accommodation of a scientific conception: toward a theory of conceptual change. *Science Education* 66 (2): 211-217.
- Rumelhart, D.E. and D.A. Norman. 1981. Analogical Processes in Learning. In *Cognitive Skills and their Acquisition*, ed. J.K. Anderson. Hillsdale, NJ: Lawrence Erlbaum.
- Séré, M.G. 1985. The Gaseous State. In *Children's Ideas in Science*, ed. R. Driver, E. Guesne and A. Tiberghien. London: Open University Press.
- Shayer, M. and P. Adey. 1981. *Towards a Science of Science Teaching*. Portsmouth, NH: Heinemann.
- Shipstone, D. 1985. Electricity in Simple Circuits. In *Children's Ideas in Science*, ed. R. Driver, E. Guesne and A. Tiberghien. London: Open University Press.
- Sjoberg, S. and S. Lie. 1981. *Ideas about force and movement among Norwegian pupils and students*. Report 81-11, Institute of Physics Report Series, University of Oslo.

- Solomon, J. 1987. Social influences on the construction of pupils' understanding of science. *Studies in Science Education* 14: 63-82.
- Solomon, J. 1983. Learning about energy: How pupils think in two domains. *European Journal of Science Education* 5(1): 49-59.
- Solomon, J. 1982. How children learn about energy, or does the first law come first? *School Science Review* 63: 45-62.
- Strauss, S. 1981. Cognitive development in school and out. *Cognition* 10: 295-300.
- Strauss, S. and R. Stavy. 1982. U-shaped Behavioural Growth: Implications for Theories of Development. In *Review of Child Development Research* Vol. 6, ed. W. Hartup. Chicago: University of Chicago Press.
- Viennot, L. 1979. Spontaneous reasoning in elementary dynamics. *European Journal of Science Education* 1(2): 205-222.
- Watts, D.M. 1982. Gravity—don't take it for granted. *Physics Education* 17(5): 116-121.
- Watts, D.M. 1983. Some alternative views of energy. *Physics Education* 18: 213-217.
- Watts, D.M. and A. Zylbersztajn. 1981. A survey of some ideas about force. *Physics Education* 16: 360-365.
- West, L. and A. Pines, eds. 1985. *Cognitive Structure and Conceptual Change*. Orlando, FL: Academic Press.
- Wiser, M. and S. Carey. 1983. When heat and temperature were one. In *Mental Models*, ed. D. Gentner and A.L. Stevens. Hillsdale, NJ: Lawrence Erlbaum.

PART FIVE: NEW APPROACHES TO SCIENCE ASSESSMENT

10. Young Children's Discussions of Science Topics

- Carey, S. 1986. Cognitive science and science education. *American Psychologist* 41(10): 1123-1130.
- Carini, P.F. 1975. *Observation and description: An alternative methodology for the investigation of human phenomena*. Grand Forks, ND: North Dakota Study Group on Evaluation. University of North Dakota Press.
- Cazden, C. 1988. *Classroom discourse: the language of teaching and learning*. Portsmouth, NH: Heinemann.
- Easley, J. 1984. Is there educative power in students' alternative frameworks—or else, what's a poor teacher to do? *Problem Solving* 6(2).

- Kanevsky, R. 1987. *Butterfly Net*. Unpublished material, developed with support of Philadelphia Renaissance in Science and Mathematics [PRISM], Philadelphia.
- Piaget, J. 1950. *The psychology of intelligence*. London: Routledge & Kegan Paul.
- Schwartz, E. 1986. An over repeating story. In *Speaking out: Teachers on teaching*, ed. Traugh, et al. Grand Forks, ND : North Dakota Study Group on Evaluation. University of North Dakota Press.
- Strieb, L. 1985. *A (Philadelphia) Teacher's Journal*. Grand Forks, ND: North Dakota Study Group on Evaluation. University of North Dakota Press.
- Vygotsky, L.S. 1962. *Thought and Language*. Cambridge, MA: MIT Press.

11. Children's Investigations of Natural Phenomena: A Source of Data for Assessment in Elementary School Science

- Arons, A.B. 1983 (Spring). Achieving wider scientific literacy. *Daedalus* 112(2): 91-122.
- Bredderman, T. 1983 (Winter). Effects of activity-based elementary science on student outcomes: a quantitative synthesis. *Review of Educational Research* 53(4): 499-518.
- Carini, P.F. 1979. *The Art of Seeing and the Visibility of the Person*. Grand Forks, ND: North Dakota Study Group on Evaluation. University of North Dakota Press.
- Champagne, A.B. 1988 (April). The Psychological Basis for a Science Teaching Model. Paper presented at the meeting of the American Educational Research Association, New Orleans.
- Department of Education and Science. 1984. *Science in Schools, Age 11: Report No. 3*. London: DES.
- Department of Education and Science. 1988 (August). *Science for Ages 5 to 16: Proposals of the Secretary of State for Education and the Secretary of State for Wales*. London: DES.
- Education Development Center (EDC). 1971. *Juba Beach, A Unit of the African Primary Science Program*. Newton, MA: Education Development Center.
- Harlen, W. and S. David. 1985. Helping children to observe. In *Primary Science: Taking the Plunge*, ed. Wynne Harlen. Portsmouth, NH: Heinemann.
- Hawkins, D. 1983 (Spring). Nature closely observed. *Daedalus* 112(2): 65-90.

- Hein, G.E. 1970. Children's science is another culture. *ESS Reader*: 87-98. Newton, MA: Education Development Center.
- Kamii, C.K. and G. DeClark. 1985. *Young Children Reinvent Arithmetic*. New York: Teachers College Press.
- Morrison, P. and P. Morrison. 1984. *Primary Science: Symbol or Substance?* New York: Workshop Center.
- Navarra, J.G. 1955. *The Development of Scientific Concepts in a Young Child: A Case Study*. Westport: Greenwood Press.
- Paley, V.G. 1986. On listening to what the children say. *Harvard Educational Review* 56(2): 122-131.
- Rowland, S. 1984. *The Enquiring Classroom*. London: The Falmer Press.
- Shymansky, J.A., W.C. Kyle, and J.M. Alport. The effects of new science curricula on student performance. *Journal of Research in Science Teaching* 20(5): 387-404.
- The Prospect Archive. 1984. *LEO*. Unpublished manuscript. North Bennington, VT: The Prospect Archive and Center for Education and Research, Inc.
- Weber, L. 1973. But is it science? In *Science in the Open Classroom*. New York: City College Workshop Center.

CONCLUSION

- Buros, O. 1977. Fifty years of testing, some reminiscences, criticisms and suggestions. *Educational Researcher* 6: 9-15.
- Carlson, S.B. 1985. *Creative Classroom Testing*. Princeton, NJ: Educational Testing Service.
- deRivera, M. 1973. Academic achievement tests and the survival of open education. *EDC News*. Newton, MA: Education Development Center.
- Edwards, D. and N. Mercer. 1987. *Common Knowledge, The Development of Understanding in the Classroom*. London: Methuen.
- Gilligan, C. 1982. *In a Different Voice*. Cambridge, MA: Harvard University Press.
- Gott, R. and P. Murphy. 1987. *Assessing Investigations at 13 and 15*. London: Assessment of Performance Unit, Department of Education and Science.
- Hein, G.E. 1968. Children's science is another culture. *Technology Review* 71(December): 3 - 11.
- Hein, G.E. 1987. The right test for hands-on learning. *Science and Children* 25(2): 8-12.

- Jones, L.V. 1989. School Achievement Trends in Mathematics and Science, and What Can Be Done to Improve Them. In *Review of Research in Education*, ed. E.Z. Rothkopf. Washington, DC: American Educational Research Association.
- Knapp, M.S., et al. 1987. *Opportunities for Strategic Investment in K-12 Science Education*. Menlo Park, CA: SRI International, NSF Contract NO SPA-85651540.
- Kohlberg, L. 1973. *Collected Papers on Moral Development and Moral Education*. Cambridge, MA: Harvard University, Moral Education Research Foundation.
- Kuhn, T. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Marshall, K. 1983. *Science: Elementary and Middle School Objectives*. Boston, MA: Boston Public Schools.
- Mullis, I.V.S. and L.B. Jenkins. 1988. *The Science Report Card, Elements of Risk and Recovery*. Princeton, NJ: Educational Testing Service.
- Murnane, R.J. and S.A. Raizen. 1988. *Improving Indicators of the Quality of Science and Mathematics Education in Grades K-12*. Washington, DC: National Academy Press.
- New York, University of the State of. 1987. *Elementary Science Syllabus*. Albany, NY: State Education Department.
- Project 2061. 1989. *Science for All Americans*. Washington, DC: American Association for the Advancement of Science.
- Raizen, S.A., et al. 1989. *Assessment in Elementary School Science Education*. Washington, DC: The National Center for Improving Science Education.

**Lesley College Elementary Science
Assessment Conference
November 4-6, 1988**

List of Participants:

Joan Boykoff Baron
Connecticut State Dept. of Education

Lynn Baum
Boston Museum of Science

Merle S. Bruno
Hampshire College

Audrey B. Champagne
(American Association for the Advancement of Science)*
SUNY - Albany

Edward Chittenden
Educational Testing Service

Richard C. Clark
Minnesota State Department of Education

Sally Crissman
Shady Hill School

Frank E. Davis
Lesley College

George Dawson
Florida State University

Rosalind Driver
The University, Leeds, England

Eleanor Duckworth
Harvard University

Hubert M. Dyasi
City College - CUNY

Brenda S. Engel
Lesley College

David Florio
National Science Foundation

Susan N. Friel
(Lesley College)*
University of North Carolina

Cille Griffith
(Randolph MA Public Schools)*

Joe Griffith
National Science Resources Center

Raymond Hannapel
National Science Foundation

Mark Hartwig
(Biological Sciences Curriculum Study)*

George E. Hein
Lesley College

Candace Julian
Technical Education Research Centers

Judith A. Kelley
University of Lowell

Dick McQueen
Multnomah, Oregon Education Service District

Jan Mokros
Technical Education Research Centers

Shelly Ornstein
Scholastic, Inc.

Jerome Pine
California Institute of Technology

Sabra Price
Lesley College

Senta A. Raizen
National Center for Improving Science Education

Susan P. Snyder
National Science Foundation

Patricia L. Stock
(University of Michigan)*
Syracuse University

Karen L. Worth
Education Development Center

* affiliation at time of conference

MONOGRAPHS BY THE NORTH DAKOTA STUDY GROUP ON EVALUATION

Coordinated by Vito Perrone

A (Philadelphia) Teacher's Journal by Lynne Strieb (\$7.50)

A Syntactic Approach to College Writing by Norton D. Kinghorn, Lester Faigley and Thomas Clemens (\$3.50)

A View of Power: Four Essays on the National Assessment of Educational Progress by Paul Olson (\$2.00)

Between Feeling and Fact by Brenda Engel (\$5.00)

Changing Schools Into Communities for Thinking by Bena Kallick (\$6.50)

Children's Journals: Further Dimensions of Assessing Language Development by Amity Buxton (\$3.50)

Children's Language and Thinking: A Report of Work-In-Progress by Edith Churchill and Joseph Petner, Jr. (\$2.00)

Critical Barriers Phenomenon in Elementary Science by Maja Apelman, David Hawkins and Philip Morrison (\$5.00)

Evaluation as Interaction in Support of Change by Ruth Anne Olson (\$3.50)

First California Conference on Educational Evaluation and Public Policy, 1976 edited by Nick Rayder (\$2.00)

Speaking Out: Teachers on Teaching by Cecelia Traugh, Rhoda Kanevsky, Anne Martin, Alice Seletsky, Karen Woolf and Lynne Strieb (\$7.50)

Teacher Curriculum Work Center: A Descriptive Study by Sharon Feiman (\$2.00)

Teachers' Seminars on Children's Thinking: A Progress Report by Bill Hull (\$2.00)

The Art of Seeing and the Visibility of the Person by Patricia F. Carini (\$5.00)

The Effect of Teaching on Teachers by Sara Freedman, Jane Jackson and Katherine Boles (\$5.00)

The Assessment of Hands-on Elementary Science Programs edited by George E. Hein (\$12.00)



PRICE: As indicated

HANDLING CHARGE: 15% for 1-10 copies; 10% for 11-30 copies; 5% for 31+ copies

CONDITIONS: All orders must be prepaid

ADDRESS: North Dakota Study Group
Box 8158
University of North Dakota
Grand Forks, ND 58202

Copyright © 1990 by George Hein, Editor

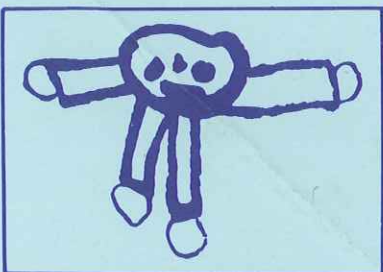
First published in 1990

All rights reserved.

North Dakota Study Group
on Evaluation, c/o Vito Perrone,
Center for Teaching & Learning
University of North Dakota
Grand Forks, ND 58202

Library of Congress Catalog
Card Number: 90-062318

Printed by the University of
North Dakota Press



A grant from the National Science Foundation
makes possible publication of this monograph

PEANUTS: Reprinted by permission of UFS, Inc.