DOCUMENT RESUME

ED 359 066 SE 053 614

AUTHOR	Meng, Elizabeth; Doran, Rodney L.
TITLE	Improving Instruction and Learning Through Evaluation: Elementary School Science.
INSTITUTION	ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, Ohio.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	May 93
CONTRACT	RI88062006
NOTE	189p.
AVAILABLE FROM	ERIC Clearinghouse for Science, Mathematics, and Environmental Education, 1929 Kenny Road, Columbus, OH 43210-1080 ($14.50).
PUB TYPE	Information Analyses - ERIC Clearinghouse Products (071) -- Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	*Educational Assessment; Elementary Education; Elementary School Science; Measurement; *Measures (Individuals); Observation; Program Evaluation; Science Curriculum; *Science Education; Science Tests; Scientific Principles; Student Evaluation; Test Construction; Testing; Tests
IDENTIFIERS	Science Process Skills

ABSTRACT

While the 1960s and 1970s came to be known as the era of curriculum development in science, it appears that the 1980s and the 1990s will be known as the time of curriculum development with strengthened attention to implementation and evaluation. This book examines the assessment of elementary school science and provides numerous examples of assessment items. Sections in Chapter 1, "Assessing Science in the Elementary School," include "Reasons for Assessing Science Learning," "Basic Kinds of Information Teachers Need," "Methods of Collecting Information for Assessment," and "Using Information to Find Answers That Fit the Original Purpose." Sections in Chapter 2, "Assessing Science Process Skills," include The Department of Processing Abilities; Assessing the Processes of Science; Using Scientific Equipment; Observing; Classifying; Using Symbols; and Predicting. Chapter 3 and Chapter 4 present detailed information on assessing concepts and problem solving. Chapter 5 is entitled "Methods of Collecting Information: How to Develop Your Own Assessment Instrument" and Chapter 6 addresses "Using the Information Gathered." (Contains 37 references.) (PR)

Improving Instruction and Learning Through Evaluation

Elementary School Science

Elizabeth Meng
Rodney L. Doran

May 1993

ERIC®

Clearinghouse for Science, Mathematics,
and Environmental Education
1929 Kenny Rd.
Columbus, OH 43210-1080

Improving Instruction and Learning Through Evaluation
Elementary School Science

Elizabeth Meng
Teachers College
Columbia University
New York, NY

and

Rodney L. Doran
University at Buffalo
Amherst, NY

May 1993

Produced by the

ERIC

Clearinghouse for Science, Mathematics
and Environmental Education
The Ohio State University
1929 Kenny Road
Columbus, OH 43210-1080

PRINTED WITH SOY INK

Cite as:

Meng, E., & Doran, R. L. (1993). *Improving instruction and learning through evaluation: Elementary school science.* Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Document development:

David L. Haury, *Editor*
Dawn D. Puglisi, *Coordinator*
Dorothy E. Myers, *U. S. Department of Education Reviewer*
J. Eric Bush and Christine Janssen, *Cover design and page layout*

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

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

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

This publication was funded by the Office of Educational Research and Improvement, U. S. Department of Education under contract no. RI-88062006. Opinions expressed in this publication do not necessarily reflect the positions or policies of OERI or the Department of Education.

ERIC and ERIC/CSMEE

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

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

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

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

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

Online Vendors:	BRS Information Technologies, 1-800-289-4277
	DIALOG Information Services, 1-800-334-2564
	OCLC (Online Computer Library Center), 1-800-848-5800
CD-ROM Vendors:	DIALOG Information Services, 1-800-334-2564
	Silver Platter Information, Inc., 1-800-343-0064

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

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

ERIC/CSMEE National Advisory Board

Table of Contents

Acknowledgments

The authors wish to acknowledge the support of Delta Inc., and the National Association for Research in Science Teaching (NARST) in the production of this monograph. We would like to thank Tom Richardson, David Butts and James Shymansky of NARST for facilitating this support and Robert Howe of The Ohio State University for initial review of the manuscript. Special thanks are extended to Stanley Helgeson for his encouragement and detailed editing and to Willard Jacobson for his continual encouragement of our professional lives. It is hoped that this publication will aid teachers and supervisors who are interested in expanding and improving their assessment efforts.

The authors would like to extend a special thanks to Francis Szucs, Jr. for preparing this manuscript, text and the numerous detailed visuals. We are indebted to many teachers, especially those in the Manhasset district, for their help in fine-tuning many of the items in this monograph.

To our spouses and families we would like to express our appreciation for encouraging us to complete this manuscript and for tolerating the many hours devoted to this task.

CHAPTER 1

ASSESSING SCIENCE IN THE ELEMENTARY SCHOOL

Introduction

For better or for worse, elementary school science has been noticed again. It is considered **IMPORTANT**. The demands of a technologically advanced country require not only the continuing replenishment of a cadre of people with sophisticated skills and knowledge but also a population capable of making rational decisions based on some knowledge of science and technology. Many educators are saying that we are not doing this; that we are losing ground; that we have to start earlier in order to educate children with the attitudes, skills, and knowledge necessary to the task. And so, elementary science education comes under scrutiny again. This time, however, there does not seem to be the interest in massive efforts to write curricula, as was the case in the 1960s and 1970s. As part of the renewed interest, the funded programs of that time[1] have been revised, modernized, and adapted to local conditions. Efforts are also underway to assist teachers in the implementation of these programs through extensive staff development projects.

Now the inevitable is beginning to happen. A variety of groups are pressing the questions: "How well are our children doing?"; "Are they becoming scientifically literate people capable of solving our science-related problems?"; "How can courses be improved to better serve our purpose of making all children scientifically literate and capable of solving problems?" While the 1960s and 1970s came to be known as the era of curriculum development in science, it appears that the 1980s and the 1990s will be known as the time of curriculum development with strengthened attention to implementation and evaluation.

As we have the responsibility for teaching the children and as we try to answer these questions for ourselves and others, another question arises. How can

[1] "Funded programs." In the 60s and 70s the federal government provided funds for the development of new science curricula for both elementary and secondary schools. Experts from science, child development, science education, and classroom teachers worked together on project teams. Out of these efforts came the "alphabet soup" programs: Science: A Process Approach (SAPA), Elementary Science Study (ESS), and Science Curriculum Improvement Study (SCIS), for example. (A later program from Britain, Sc 5-13, will also be mentioned in this monograph.)

we gather the information necessary to evaluate this kind of learning, the kind that will reflect the knowledge, skills, and attitudes recommended as being important to our children's progress? An analysis of the task confronting us yields a structure within which we will examine assessment of elementary school science.

1. What are the *reasons* for which we are assessing science achievement? Who is asking the questions? What are the questions?
2. What kinds of *information* are needed to answer the questions?
3. What *methods* or ways of collecting information are most useful and appropriate?
4. How can we *use the information* to find answers to the original questions?

Reasons for Assessing Science Learning

There are many reasons that teachers need information about children and their academic programs. Because of this, the kinds of information required are not the same for each reason. In addition, the ways of gathering it may, of necessity, also be different. It is important to keep this thought in mind as we discuss assessment in elementary science. Contrasting examples may help to illustrate this point.

During the normal operation of a classroom, teachers gather a considerable amount of data about children's knowledge and abilities. This is often done informally and stored in our heads. We use this information to plan next steps in our program based on what children can do, what they need help in, and what the curriculum requires of us. The kind of information gathered for this type of purpose is of a very specific, detailed nature that is limited in scope. For example, can the children set up a simple circuit to light a bulb? If a wire is missing, do the children know that one is needed in that particular spot? Do the children use their observations to infer the difference between series and parallel wiring? Based on the answers to these questions, we make decisions on whether or not to present additional activities and what kind; whether or not to go on to the next concept; and how we should pace future work to maintain motivation and interest.

In contrast to this specific, informally gathered data, is the kind of information that a board of education might want about the science achievement of elementary students in the district. While they might also want strengths and differences identified, their primary concern probably would be about how the students compare to those in neighboring communities, the state, or the nation. The information for this purpose would be gathered by more formal means, such as tests, and would be less detailed and specific. More aspects of the curriculum would be assessed and the results would be evaluated or judged against a set of criteria or norms. The content of the assessment, the method of administering it, and ways of scoring it would all have to be standardized in order to make reliable comparisons between individuals and groups. This latter type of assessment, which often results in one general score or a few subscores, would be of little help in making decisions about what to do *tomorrow*.

Some specific situations are described below in which the purposes for assessing differ.

1. *Day-to-Day Classroom Planning and Teaching* This usually involves detailed, specific information limited to those aspects of the activity or curriculum being covered during a time interval. It may not involve the same information for all children and is often an ongoing process designed to interfere as little as possible with instruction.

2. *Assessment for Record keeping and Report on Progress* For this purpose, a broader range of information is necessary which includes a balance of the major objectives of the curriculum. The resulting records are in large, general categories but will be specific enough to identify rate of progress. Criteria or "bench marks" are often used denoting points along a continuum of development. Categories for recording must be applicable over time and planned in order to gather appropriate baseline data for comparison; such as, "what a child knows and could do." All children may not be assessed for all categories on any given occasion or by any one method of assessment.

3. *Reporting to Parents and Others Concerned with a Child's Progress* This mode is defined in even more general terms than either of the two types above, but still reflects a balance of curriculum objectives. It is usually given in clear, unambiguous terms and includes an evaluation of progress and/or a comparison to some norm. The criteria used for making judgments must be clearly described and applied similarly to all children. The methods of assessment must be appropriate for gathering information over time and for repeated use so that results can be compared.

4. *Assessment for Curriculum and Program Evaluation* This last example may not include all aspects of program, but only those in question or considered to be of primary importance. Items are not tied to specific activities, but to a basic set of goals of science instruction. No one child is assessed over the total range of objectives, but sampling must be representative of the population and large enough to be reliable. This mode usually involves multi-format assessment, such as a combination of written tests, practical tests, observations, etc. The evaluation is accomplished in terms of a set of criteria or in terms of strengths and weaknesses, as evidenced by student performance.

Basic Kinds of Information Teachers Need

What a curriculum consists of and how it is translated into a program is determined by our beliefs about what science education should be in relation to how children learn best. Beliefs about science education also change as we learn more about the way children learn and as society's needs and interests change. This creates identifiable characteristics of programs which most science educators seem to agree upon at any particular time. Most of the present goals in science education are evident in the following statement (Figure 1.1) of policy used to guide the development and implementation of a school program (Manhasset Public Schools, 1985). We are including it here in order to illustrate the types of programs that current efforts in assessment are focusing on.

It is fairly clear that any program developed from this policy statement would include concepts (the big ideas of science), science processes, and scientific attitudes. It would also be based on the active exploration of objects and phenomena and often focus on a problem to be solved. Assessment would need to be undertaken within these parameters.

1. *Process* Process skills are used to make sense of the world and our experience as part of it. If science is a search for patterns, then the process skills are the tools of that search. Some, such as observing, looking for patterns in observations, and recording, are common to many curriculum areas. Others, such as identifying and controlling variables, planning, and carrying out a fair test or investigating are more specific to the field of science. Through natural maturation and by continual use in a wide variety of relevant situations, a child hones these skills and becomes better able to recognize the situations in which they are appropriately used. There has been renewed attention given to the area of process development, as it is recognized that both scientists and informed citizens need to learn how to effectively process information in order to solve many of our science-related world problems. Process skills and their assessment will be discussed further in Chapter Two.

2. *Content* While the process skills are the tools used in the search for patterns and solutions, the content area deals with the patterns that are found; the big ideas we construct out of the bits and pieces of our experiences. While at first closely tied to the concrete, actual experience and the data collected from it; the ideas, or concepts, gradually become refined and changed through maturation and experience into more generalized ideas. These generalizations are further refined into sophisticated abstractions of a universal nature. Some of the big ideas that elementary science programs focus on are: cause and effect in simple change; effect of forces; the properties of common materials (in the physical sciences); and concepts of organisms, life cycles, and ecological systems (in the biological sciences).

Science in the Elementary School

Science is basically a search for patterns. Science education on the elementary school level involves putting children in situations that allow them to search and discover patterns and then to apply these patterns to new situations. The child's developmental level, background knowledge, skills, and interest will influence what is chosen to be studied and how it is presented. The following statements guide us in developing and implementing science programs on the elementary level (K-6):

Children should be encouraged to be curious about their environment - both physical and biological. They should be encouraged to ask questions and to wonder. A curious, enthusiastic teacher is an important component.

Experience with appropriate hands-on activities is essential to the development and retention of the "big ideas" of science (patterns of science). It is not enough to read about science.

Children should be exposed to the idea of a "fair test." With teacher guidance even young children can determine "fairness" if few variables and concrete, familiar experiences are used.

Children should have experience collecting, organizing, and communicating data gathered in a real situation. They should be given the opportunity of applying the resulting generalizations to a new situation. In doing so, they will develop facility in the use of the process skills of science.

In elementary science education teaching is not telling. The teacher assists the students in organizing their thinking and asks the right questions so that children can learn to evaluate the effectiveness of their own work.

Science experiences should be "whole." Where appropriate, the other subject areas should be included so that children understand the topic in the total context.

Recording and communicating in a variety of ways helps the child clarify and find meaning in science experiences.

The use of measuring and other tools from mathematics helps the child discover relationships and patterns. While the youngest children first have qualitative experiences (exploring natural phenomena and looking for similarities and differences), they are soon able to compare using the mathematical tools. This also supports the mathematics program

Many current problems in our society need an understanding of science concepts before a solution can be found.

Science is not a difficult subject. All children have the right to be scientifically literate and to enjoy science experiences.

The quality of teacher guidance is the most important ingredient of an elementary science program.

Figure 1.1. Science in the Elementary School

Unfortunately, some facets of content, such as names, definitions, and facts, are more easily assessed than other components of equal or greater importance and tend to dominate many teacher-made tests. This results in an assessment of limited scope and sends a message to the student about what the teacher "really" thinks is important. (It may also distort the intention of the program.)

In addition to knowing and applying concepts to familiar situations, the big ideas are used in the creation of new knowledge in different situations. Assessment should, therefore, be multi-faceted in terms of how the concepts are used and generated, as well as in terms of the development of the concept. Assessment of content outcomes will be discussed in Chapter Three.

3. *Attitudes* Much of what has been said about development of processes and concepts also applies to the area of attitudes. They develop and mature over time, change as they develop, and are often difficult to assess. Some are part of the general instructional program, such as curiosity, cooperation, and perseverance, and others are more specific to science; for example, confidence in solving problems and respect for living things. Attitudes affect performance in the other areas as well as being affected by them. Recent efforts to develop methods of assessing attitudes seem to hold some promise. This is an important part of science education but not within the scope of this monograph. Therefore it will not be included.

4. *Problem-Solving* Problem-solving is considered by many to be a product of science education. To others it is a means, or method, involving certain steps; such as planning an investigation, obtaining the data, and formulating conclusions. We consider problem-solving, or inquiry, to be the intersection of three sets: attitudes, process, and concepts. This is where a child puts everything together and deals with an actual, total situation. It is a creative, complex operation that is probably much more than the sum of the individual process skills. For example, a problem has to be recognized as existing and then must be translated into a question that can be investigated or set up as an experiment. It is also not enough to just observe. A judgment must be made about what and when to observe, and relevant observations must be sorted out from those that are irrelevant to the problem. A Fair Test or controlled experiment must be planned and performed and the conclusions drawn based only on the data that was collected—and so on. Decisions also have to be made on how to organize and record the data. It is not enough to know how to use the tools of science. Knowing when they are to be used and in what relation to other tools is part of the problem-solving process. As attention turns to the development of critical thinking skills in the total curriculum, problem-solving in

science also becomes increasingly important. While a few of the basic process skills of science that are prerequisites for effective problem-solving can be assessed separately, most of the complex and interacting skills of problem-solving must be assessed within the total problem-solving situation, which includes skills, concepts, and attitudes. This will be discussed further in Chapter Four.

Every one of the reasons for assessing has one thing in common with the others: a need for a description of where individuals or groups of children are in their development. What levels of proficiency have been attained in the areas of process, content and attitudes? How well do they solve problems? Growth in these areas involved a continual, progressive change affected by a child's experience and maturation. Therefore, it is not enough to look for evidence of a process, concept, or attitude when making plans and judgments concerning children's science learning. Knowing the stage of development, at a particular point in time, of the various components is an absolute necessity for effective assessment—and teaching.

Methods of Collecting Information for Assessment

There are several methods of collecting information for assessment purposes. Most of them are familiar to both students and teachers. For example, data can be collected using written tests (paper and pencil), practical tests (also called lab tests, authentic tests, or performance tests), observations, and interviews, or through the analysis of some product such as a project. All are useful and appropriate for *some* of the objectives of science in the elementary school. None are both appropriate and useful for all objectives. Chapter 5 will present practical suggestions for developing assessment instruments of several formats.

The most widely used method of collecting information on classroom learning is the written, or paper and pencil, test. These tests are extremely useful for assessing student achievement on content objectives. On the other hand, "practical" exams require students to observe objects or phenomena, manipulate equipment and materials, measure or estimate, record and organize information, plan an investigation and implement it, as well as explain and interpret the data collected. While practical assessment can be used to assess content, it is more often used for process and/or problem-solving objectives.

Valuable information can also be collected by teachers observing individual students, small groups, or an entire class. Much of our informal information gathering is done this way or through interviews and discussions with the children. To be effective, some focusing or structure is necessary, as well as a planned method for recording the information gathered. Such techniques can focus on process and problem-solving skills, as well as on some of the content objectives.

The analysis of student products, such as projects and reports, has been widely used by teachers to determine students' knowledge and skills, and their ability to plan, conduct, and report on investigations. This category also includes the "free" writing of young children, as well as drawings, diagrams, charts, and other forms of expression of students' interaction with science experiences. This method also requires a structure and focusing, as well as a system for recording the information gathered, if it is to provide information for assessment and not just a vague "grade."

Any of the above techniques may be used in conjunction with another. For example, a discussion with a student can serve to clarify the meaning of an action or a written response. An observation of a child involved in practical work can indicate the extent to which a child *will*, as well as can, use a skill.

There is another dimension to the method of gathering appropriate and useful information besides format choice. The focus of educational evaluation is often described as being within the three domains of educational objectives, as initially described by Bloom and his colleagues (Bloom, et. al., 1956). They are the cognitive, the psychomotor, and the affective domains.

The cognitive taxonomy has had a tremendous impact on curriculum development and educational research, as well as on assessment. In this volume we will use a slightly modified scheme with three categories: Knowing, Using, and Extending. The authors believe that this compressed variation is a good match for the objectives in most elementary science programs, and one that can be used relatively easily. The link to the six Bloom categories is shown below.

a.	Knowing	knowledge
b.	Using	comprehension, application
c.	Extending	analysis, synthesis, evaluation.

Test items using any format are usually designed according to one of these levels of cognitive outcomes. For example, a student might be asked a question to determine whether he *knows* a concept, whether he can *apply* that concept, or whether he can *extend* his knowledge about electricity by employing the concept to analyze a situation. As many of the objectives of elementary science programs emphasize more than *knowing* a concept, tasks used for assessing cognitive objectives should also require the student to demonstrate more than a knowledge of concepts (or facts).

Although the Bloom team never delineated levels for the psychomotor domain, it can be considered to include many of the objectives and abilities related to observation and practical/laboratory work. Active student involvement is central to the view of science instruction portrayed by most current science curriculum projects, state and local guides, and some commercial textbooks. The evaluation of these objectives has lagged behind the work in the other domains. Recently, several assessment efforts have been initiated in the practical or laboratory area of elementary science. This volume will attempt to present guidelines for the developing and administering of assessment instruments in practical, paper and pencil, and other formats, in Chapter 5.

Using Information to Find Answers That Fit the Original Purpose

Once an assessment has been completed, the information gathered is generally used to make some sort of judgment or evaluation. The basic reasons for assessing student achievement, given at the beginning of this chapter, could be reconceptualized in terms of three dimensions of evaluation; diagnostic, formative, and summative.

The purpose of diagnostic evaluation is to determine *before*, or at the beginning, of instruction, what the student possesses in terms of previously acquired background experiences, skills, attitudes, and misconceptions. This would indicate which students need to receive special help to complete missed areas or to develop necessary skills for the next unit of instruction. It would also identify students who already possess the intended level of skill and/or concept development.

Evaluation for a different purpose occurs at some point *during* a segment of instruction. This is termed "formative" evaluation. It helps teachers to determine the degree to which students are learning the intended information or mastering the planned skills. It provides feedback to both student and teacher as to whether they are "on schedule." Such information is not intended for grading purposes but to help the teacher to adjust the rate of instruction, assign remedial activities, and plan alternative experiences.

Summative evaluation occurs at or *near the end* of instruction. It is what most of us think of as "testing"—assessment for the purpose of assigning grades, determining placement, and identifying progress. This certainly is an important part of evaluation, but it is not the whole picture. Because this type of evaluation is perceived as being so important by parents, teachers, administrators, guidance counselors, etc., we must be very careful to provide balanced, fair, and valid assessment.

Any judgment or evaluation is done against a "yardstick," or frame of reference. Historically, most of the efforts in educational evaluation have been what is called "norm referenced." Individual students, classes, or schools are compared to some norming group: commonly, the district, the state, or the nation. Such statements reflect a comparative mode in which there are "winners" and "losers." In many contexts such a frame of reference and comparisons are valid and appropriate. However, it is a requirement of such a system that one-half of each cohort group (students, classes, or schools) will be labelled "below average."

In other cases, a different rationale is presented, primarily within the "cooperative" teaching/learning mode. In these cases the teachers and students, working together, strive to learn as much of the information and skills as possible. A "criterion-referenced" system has evolved as an attempt to facilitate evaluation within this framework. A basic requirement is the formation of *a priori* statements in quite specific terminology about the desired outcomes. Each student, class, or school is then judged as being successful once they have demonstrated that they have achieved, or mastered, the stated objectives or outcomes (criteria). It is certainly possible, indeed desirable, that a majority of the group "pass" or "master" a given unit of instruction.

Information intended for diagnostic, formative, or summative evaluation could be presented in a framework of either norm- or criterion-referenced evaluation. Historically, summative evaluation has used a norm-referenced yardstick; however, recent efforts to evaluate programs have used a criterion-referenced approach. Thus, it is possible to infer strengths and weaknesses in the program and make necessary adjustments in the future. Useful information would be yielded for diagnostic and formative purposes if a criterion-referenced system were also used. More purposes are served by the criterion-referenced system. Chapter 6 will focus on the uses of information collected for description and evaluation of students and their development as well as the instructional programs.

This chapter has served as an introduction to ideas that will be developed in the remaining chapters. A general overview has been given for the reasons, or purposes, for assessing science in the elementary school. The rest of the volume will address the kinds of information gathered for assessment, the methods of gathering data, and the appropriate use of the data in light of the original purpose. Illustration of items that have been used will be included, along with tips on how to develop your own assessment instruments.

CHAPTER 2

ASSESSING SCIENCE PROCESS SKILLS

If science is a search for patterns, then the searching skills should assume an equally important position in science education as the patterns found through this searching. Although we have consistently focused on the knowledge (e.g., facts and patterns) of science in science education, the programs of the 60s also put children in the role of the scientist. Lists of what were considered the intellectual and manual skills of "doing" science were used in the planning of science experiences for children. For example, Science: A Process Approach (AAAS, 1968), the first of the major programs, developed a set consisting of two levels of skills; the primary process skills, for younger children; and the integrated process skills, intended for older and more experienced students (see Figure 2.1). The second level applies more directly to the complex activity of carrying on an investigation or problem-solving (see Chapter 4). SAPA was developed with support from the National Science Foundation. Its development was coordinated by the Committee on Science Education of the American Association for the Advancement of Science (AAAS). Other lists of skills were developed by schools, states, and textbook authors. Some authors separated and others grouped the various processes and occasionally assigned somewhat different names to the categories. Whatever names the skills were given, and however they were organized, they were all considered illustrative of how scientists work, and therefore of what children should be taught.

Science Process Skills — developed for SAPA	
Primary Skills	**Integrated Skills**
observing	making operational definitions
recognizing time/space relationships	controlling and manipulating variables
recognizing number relationships	formulating hypotheses
classifying	experimenting
measuring	interpreting data
predicting	
inferring	

Figure 2.1. Science Process Skills—developed for SAPA

The Development of Processing Abilities

While the various "laundry lists" of skills may have been helpful in focusing attention on this important area of science education, they contributed little to our understanding of the development of processing abilities. As mental processes mature and the ability to investigate develops, how children think and what they can do changes dramatically. For example, a child's ability to classify depends both on the maturation of the mental/intellectual structures and on the variety of experiences in classifying. Very young children often forget the attribute with which they started classifying and switch as they go along, apparently without noticing the inconsistency. Later, they are able to complete the task of grouping two sets of objects based on mutually exclusive categories, as long as they are dealing with concrete materials, but often cannot name the attributes of each set. Gradually, the ability to construct complex classification systems based on abstractions, not observable attributes, develops. Thus, the experiences of classifying in ever increasing degrees of sophistication, plus the normal maturation of the cognitive structures, lead to mastery of the skill of classification.

While implying and assuming development and change, most elementary school science programs have not described levels or stages of progress in terms of observable behaviors that are sufficiently specific for classroom use or, therefore, for productive assessment. For example, those listed for ESS describe what a child will do for a goal, such as measuring. "The student will demonstrate the ability to measure length, area, volume, weight, temperature, force, and speed" (Aho, et. al., 1974). While helpful, the phrase "demonstrate the ability to..." is not defined in terms adequate for assessing varying levels of ability to measure. (Neither does "develop skill in..." suit this purpose.) Specific descriptors or "benchmarks" are needed for both specific levels and for a continuum that denotes increasing mastery or competence.

Recent work, done in the area of process mastery, is of special interest to anyone wishing to determine at what stage children are in their developing scientific abilities. Each of the three schemes, which will be used for purposes of illustration, were derived from theories of child development and current goals of science education. Their purposes and frames of reference vary somewhat. Some are more descriptive of how children think, others are more illustrative of observable behaviors. The number of categories also varies, as does the number of steps on the continuum. All three provide benchmarks along a developmental continuum which reflects an interaction between both maturation and experiences.

The description of categorization (classification) in the first scheme (Shayer & Adey, 1981), shown in Figure 2.2, clearly focuses on how a child thinks. Benchmarks are provided for the levels of cognitive development (pre-operational through formal) as defined in Piagetian terms. On the lower developmental levels, the ability to classify is clearly tied to simple "concrete" situations (the actual objects), becoming more generalized later and finally progressing to a complex "abstract" nature (the universal laws and theories of science). The set on measurement, while being more descriptive of observable behaviors, still uses this frame of reference and links the planning of appropriate experiences and assessment to a child's cognitive development.

Skills Classified by Developmental Levels

	Type of Categorization or Classification
1 pre-operational	Thought is associational, and association of one aspect (e.g., height) not linked to another aspect (e.g., breadth) on anything but an immediate perceptual or temporary basis. Thus has difficulty classifying objects into even two groups as successive judgments on one object are contradictory.
2A early concrete	Elementary classification. Sets of objects are classified according to one major criterion at a time, e.g., color, size, shape, etc. Children can switch criteria. Soon they can also multiply classifications, e.g., "big-blue-squares/small-blue-squares," "red-big-squares/ small-red-squares."
2B late concrete	Class inclusion and hierarchical classification. Classification is still the dominant mode of categorizing reality, but now the classes are less tied to one simple property, and can also be partially ordered, e.g., "animals — flying animals — domestic birds." Bi-polar classifications such as "Acids and Alkalis as opposites" are possible.
3A early formal	Generalization. Now the classifying operation is used to impose meaning over a wide range of phenomena. A general formula like $V = h \cdot l \cdot b$ will be used as an instruction for computing volume. Asked to choose the next term in the series, "Etna—volcano—....," this student would pick "mountain."
3B late formal	Abstraction. By contrast, would prefer "geological notion" to "mountain" as next stage of categorization. Because of the multivariate nature of reality it is sometimes more powerful to search among the many properties for the essence of the underlying association. "Mountain" is a class, but "geological notion" abstracts so that connections with non-mountains can be explored. In $V = h \cdot l \cdot b$, the way in which h and b vary in relation to one another for constant l and V.

	Skills in Measuring and Interpreting Relationships
2A early concrete	Makes measurements by comparing beginning and ending of object/journey with rule in simple whole numbers.
2B late concrete	Bar diagrams, histograms, idea of *mean* as the center of a histogram, and variation as its breadth. Graphical relationships of first order equations. Interpretation of graphs where there is a 1:1 correspondence with the object modelled, e.g., height/time relationship for the growth of a plant.
3A early formal	Interpretation of higher order graphical relations, and use of problem-solving algorithms, e.g., $P_1V_1 = P_2V_2$ for gas pressure calculations. Can make interpretations which involve relations between variables in a graph, e.g., in a distance/time graph will see that a horizontal section means "standing still" and that a vertical section is impossible.
3B late formal	Interpretation of higher order graphical relations in terms of *rates* (instantaneous slopes) and reciprocal relationships; conceptualization of relationships between variables, e.g., in $V = hlb$, if l rises (V constant), b and/or h must drop proportionally.

Figure 2.2. Skills Classified by Developmental Levels

The next scheme, illustrated in Figure 2.3 (The Wisconsin Department of Public Instruction, 1970), provides a list of measuring tasks that are roughly associated with a child's developing abilities and skills. Although split into more levels and composed of observable behaviors, it is not necessarily "better" than the others described here. The purpose and other constraints would, of course, affect any choice. These might include such factors as how much and what kind of information is needed (as well as what is appropriate) in order that the assessment be both effective and manageable. For example, a grade level assessment of one skill involving a rather homogeneous group of children might concentrate on one level/section of the scale. If many skills are to be assessed, items constructed for a few benchmarks would be more manageable. How the information is to be gathered (by observation, written test, etc.) would also influence the number and kinds of benchmarks preferred.

More categories are included for illustration from the next and final scheme (Harlen, Darwin & Murphy, 1978) because it was specifically designed for assessment purposes in a variety of classroom contexts (see Figures 2.4, 2.5, and

A Process Sequence for Measuring

a. Ordering objects by inspection in terms of magnitude of selected common properties, such as linear dimension, area, volume, or weight.

b. Ordering objects in terms of magnitude of properties by using measuring devices without regard for quantitative units.

c. Comparing quantities, such as length, area, volume, and weight, to arbitrary units. Comparing time to units developed from periodic motions.

d. Using standard units for measurement.

e. Selecting one system of units for all related measurements.

f. Identifying measurable physical quantities which can be used in precise description of phenomena.

g. Measuring quantities which depend upon more than one variable.

h. Converting from one system of units to another.

i. Using and devising indirect means to measure quantities.

j. Using methods of estimation to measure quantities.

Figure 2.3. A Process Sequence for Measuring

2.6). It too is based on a cognitive theory, with the categories divided into two groups: one for younger children and the other more appropriate for older students. Where the same category is often used for assessment on both levels (younger and older), there is some overlap evident in the descriptions provided as benchmarks. For example, see Figure 2.4, "Observing." Within each group there are five possible benchmarks or levels; the three with detailed descriptions and in-between or transition positions. After a period of using a system such as this, the benchmarks become important operational definitions of student achievement in skill areas.

The format provided in this scheme makes it suitable for a record of a child's progress. The categories chosen from this scheme are of interest to those who would want to assess the types of specific skills listed at the beginning of this chapter. Figure 2.4, describes levels of the observing behaviors for both younger and older children. Figure 2.5 was chosen to illustrate the development of a key skill (classifying) for assessment of younger students' learning. Lastly, Figure 2.6 illustrates a more complex skill, "Finding Patterns in Observations," intended for older students.

Levels of Mastery — Observing Skills for Younger Children

Rarely gives any indication of noticing new or unusual things unless they are pointed out to him.	Shows that he notices some of the new or unusual aspects but misses details which can be observed, making limited use of his senses.	Uses several senses where appropriate, noticing sequence and a variety of details with reasonable accuracy and objectivity.

Levels of Mastery — Observing Skills for Older Children

Makes limited use of his senses, noticing only some of the things which can be observed in the situation or only those which are pointed out.	Makes all kinds of observations using several senses, though not able to discriminate the more important from the less important observations for the enquiry at hand.	Makes wide-ranging observations and can select from them the information relevant to a particular problem or enquiry.

Figure 2.4. Levels of Mastery—Observing Skills (Part 1)

Levels of Mastery — Classification Skills for Younger Children

When he groups objects together there is no consistency between the objects in the groups and his declared intentions in grouping.	Can consistently separate things which have a chosen feature from those which do not, but cannot then select a different feature and regroup the objects according to that feature.	Having sorted objects into groups according to one feature, can re-sort the same objects according to a different feature, which he selects himself.

Figure 2.5. Levels of Mastery—Classification Skills

Levels of Mastery — Finding Patterns In Observations with Older Children

Does not related findings to the purpose of the enquiry or notice any patterns there are to be found without considerable help.	Attempts to look for patterns in findings but rarely suggests possible explanations.	Makes reasonable inferences which fit the evidence and makes some attempt to explain the patterns which he finds in his observations.

Figure 2.6. Levels of Mastery– Finding Patterns

The reader will want to compare the schemes as conceived by the different authors (or teams). For example, the Shayer scheme is organized around stages of cognitive development with thought patterns, or behaviors, of each level specifically identified. The two other illustrations (Wisconsin & Harlen, et. al.) emphasize observable behaviors, with the Wisconsin being broken down into more discrete steps. Neither of these last two specifically identifies characteristic behaviors according to the various Piagetian stages, (although the steps given are progressive). These differences, and perhaps others, might be determining factors in choosing one or the other to help when developing one's own version for specific assessment purposes.

Assessing the Processes of Science

Some skills can be assessed directly and are considered important to all areas of the curriculum, not just science. Without them, data could not be collected and organized for use in problem-solving. For example, a child wishing to determine which of four materials is the best insulator would have to read a thermometer and record information in some organized form, such as a table, and then graph the data in an appropriate manner. S/he would also have to be able to interpret the graph in order to find a pattern and come to any conclusions. These skills are commonly thought of as tools of mathematics and are also used in the social studies area. Some of the other skills that are more specifically used in science interact and frequently depend on the use of previously acquired science information. The interaction of these skills and knowledge results in the ability to plan and perform part or all of a science investigation and to apply science concepts to new situations. These latter types of skills, which are usually tested as a group within the actual context of an investigation will be discussed in Chapter 4, "Problem Solving." Those that are often assessed separately will be presented in this chapter.

Various assessment tools have been used to monitor the development/ mastery of a variety of specific skills. There may be a question about the validity of such an assessment, for skills are not as discrete as a list might seem to imply (e.g., primary process skills of SAPA). Therefore, it must be cautioned that there is a strong possibility that one may be assessing different skills when testing separately, as opposed to within a problem-solving situation. However, there may be times when the assessment of a specific skill is desirable (e.g., for diagnostic purposes).

A variety of methods have been used to test for the acquisition of a specific skill, such as paper and pencil tests, observation of performance with materials, and various adaptations of both. Each has both advantages and disadvantages. For example, the use of a totally paper and pencil test would allow for a greater number of skills to be assessed in a short period of time. Answers could be graded by hand, or machine, in a short time by relatively inexperienced people. This type of test tends to be inexpensive, especially if the student does not write in the test booklet. On the other hand, it depends on the child's ability to read, visualize, and comprehend what is being asked. There is usually no provision for the child to seek clarification of the task. The relevant variables are often given in either the stem or options, so the intellectual task is limited. (For example, the student may be cued to a limited number of attributes, the names of which have been given by the test maker.) Actual materials are probably not included, even though most students are probably concrete operational and would benefit from their use. A certain degree of skill in writing would be necessary if the students were expected to supply the answer, not choose one.

Observation of students involved in using a particular skill, and discussion with them, also has advantages and disadvantages. Most of them are the flip-side of the factors discussed for paper and pencil testing in the preceding paragraph.

For example, young children are better able to demonstrate skills and understandings than they are able to tell about them. Although this format could yield the most valid and reliable information about student skill development, it can be time consuming and expensive. Therefore, a mix of the two formats is often used. For example teacher questions are often scripted in an effort to standardize the test situation. A test booklet is used for the student to record data and conclusions, etc. The items in the test booklet may require the use of manipulative materials. The options are not given (as in multiple choice); the student supplies the answer. Diagrams are used to present as much information to the student as possible, and what reading is required is kept to a limit. The language used is also kept simple and to a minimum so that another variable is not hidden in the task. (A further discussion of format choice is found in Chapter 4, "Problem Solving.")

Some items that have been developed to assess specific skills have been chosen for illustration, with the emphasis in choice being placed on the skills regarded as being shared by many discipline areas of the curriculum. An attempt was made, in choosing the items, to present a selection intended for different purposes, different group sizes, and various age groups. Items will be illustrated that assess the following skills; using scientific equipment, observing, classifying, and using symbols.

Using Scientific Equipment

The first illustration, "Making Comparisons Using a Balance" (see Figure 2.7), includes the instructional objectives for the learning activity in addition to the "appraisal" item (AAAS, 1968). It is intended to gather information for diagnostic purposes from the group via student demonstration of appropriate behaviors using an equal arm balance. The teacher interacts with the group, and children take turns in this group setting. If the group seems to have mastered the skill at the level indicated in the objective, the teacher would plan to move on to the next objective in the hierarchy. If not, other activities would be chosen to give those needing additional experience the opportunity to master the skill before proceeding. Note that most of the assessment from the SAPA program is based on what the child "does," not on what is said or written.

The following illustration is part of an assessment item on the process test used for the Second International Science Study (SISS, 1986). Unlike the previous item (from SAPA), the student is expected to read and follow directions and write an answer in the Student Test Booklet. Skill in reading a thermometer is requisite to completing the rest of the task, which consists of predicting the temperature of the mixture after having discovered the pattern involved. There are problems grading a student's accuracy in reading a thermometer, unless room temperature water is used or the tester checks the temperature right after the student. Other skills, such as weighing or measuring, can be tested more easily by using an object of known dimensions or mass.

Making Comparisons Using a Balance

Objectives

At the end of these activities the children should be able to:

1. Order objects whose weights differ appreciably by lifting them.

2. Demonstrate how to balance the force on one end of an equal-arm balance with a force on the other end.

3. Demonstrate how to weigh a small object with an equal-arm balance by counting out such weights as paper clips, marbles, or small blocks of wood.

4. Demonstrate how to find out how much heavier one object is than another with an equal-arm balance.

Appraisal

Show the children three objects that are of about the same weight, such as a wooden block, a box of paper clips, and a bag of crayons. Ask three children in succession to order these objects from the lightest to the heaviest. If the children disagree in their ordering, ask them how they might do a more accurate job of ordering these objects by weight.

1. Do they suggest the use of the equal-arm balance?

   Have three children order these objects from the lightest to the heaviest using the equal-arm balance.

2. Do the children compare the weights of the objects by successively placing them on opposite ends of the equal-arm balance? Do the three children agree in their ordering of the objects?

   Place the lightest object on one end of the balance and the heaviest on the other and ask one child to show you how the heaviest object would be balanced.

3. Does the child balance the heavier object by applying additional force on the side of the lighter object?

   Place a box of paper clips and some marbles on the table. Ask a child to find out how much heavier the heaviest object is than the lightest object.

4. Does the child add these small objects to the side of the lighter object to see how many are needed to balance the heavier object?

Figure 2.7. Making Comparisons Using a Balance

Predicting Temperature of a Mixture

Use the water in the cups marked "X" and "Y" to test the changes in temperature when you mix hot and cold water.

1. What is the temperature of the water in cup "X"?

 ____

2. What is the temperature of the water in cup "Y"?

 ____

3. What do you think will be the temperature of the mixture when water from cup "X" and cup "Y" are poured into the larger cup?

 ____

Now pour the water from cup "X" and cup "Y" into the larger cup. Stir the mixture.

4. What is the temperature of the mixture?

 ____

5. If the temperature of the mixture is different from what you predicted in Question 3, what might be the explanation?

 ____

 ____

 ____

6. What do you predict would be the temperature if you mixed equal amounts of water at 5°C and 75°C? Explain.

 ____

 ____

 ____

 ____

Figure 2.8. Predicting Temperature of a Mixture.

The remaining illustrations used to assess skills of working with simple laboratory equipment, are primarily paper and pencil tasks, with answers being chosen from a set of limited options. A diagram is usually given in lieu of the actual equipment. Sometimes this skill is a required component of an item. In this case, the tester would want to know if the student could use the skill adequately enough to allow completing the task which also required another skill or some specific concept. Sometimes it is a straight-forward item just to test that one skill. Two exclusively paper and pencil items are given below. Figure 2.9 is from the SISS achievement test (SISS, 1984). Figure 2.10 is an item on the Jefferson County Elementary Science Test (Jefferson County Schools, 1982). This last item may only require skill in reading a number line.

Example of Measurement Item with Diagram

How long is the block of wood shown in the diagram?

A. 10 cm
B. 20 cm
C. 25 cm
D. 30 cm
E. 35 cm

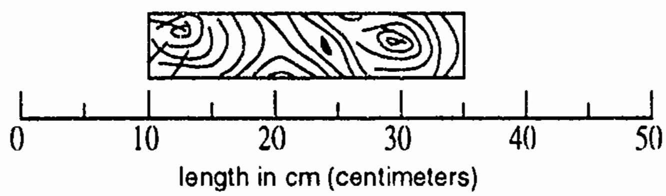
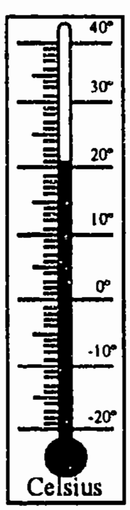


Figure 2.9. Example of Measurement Item (Item 1)

Example of Measurement Item with Diagram

Larry wanted to know the temperature outside. His thermometer looked like this:



What was the temperature outside?

A. -21° Celsius
B. 19° Celsius
C. 21° Celsius
D. 23° Celsius

Figure 2.10. Example of Measurement Item (Item 2)

Observing

An item from SAPA Level A (AAAS, 1968) (see Figure 2.11) was developed to assess the skill of observation. The objectives are as follows:

1. Identify and name the changes which occur when a solid changes to a liquid, including changes in properties such as height, width, color, temperature, and shape.
2. Distinguish between solid objects which melt and those which do not melt, under specific conditions.

The activity involves observing solids changing to liquids. This SAPA competency measure, while intended for individuals, could be given to all the children at the same time. (This, of course, would depend on their writing ability.) There are scripted questions for the teacher to ask. Children do not manipulate the materials themselves but instead watch a teacher demonstrate the activity.

Competency Measure for Observing

COMPETENCY MEASURE

(Individual score sheets for each pupil are in the Teacher Drawer.)

Put a piece of rock salt in one custard cup, a small piece of beeswax in a second, a piece of wood in a third, and a marble in the fourth.

TASK 1 (OBJECTIVE 1): Ask, "Each of these objects is in a dish. In what other way are they alike?" An acceptable response includes the statement that the objects are all solids.

TASK 2 (OBJECTIVE 1): Tell the child, "I am going to pour hot water over each of these substances. Watch carefully for any changes which you might see." Pour hot water into each cup until it is about half full. Point to the cup which had the rock salt in it. Ask, "What changes took place in this solid?" An acceptable response includes the observation that the salt changed from a solid to a liquid. It must include the word "liquid."

TASK 3 (OBJECTIVE 1): Repeat task two, pointing to the beeswax. An acceptable response includes the observation that this solid changed its shape, or that it turned from a solid to a liquid, or melted.

TASK 4 (OBJECTIVE 2): Repeat task two, pointing to the piece of wood. An acceptable response includes the observation that the wood is floating on the water, or that there is a change in the position of the wood in the dish, or that the wood has not changed its shape.

Figure 2.11. Competency Measure for Observing

The skill of observing can also be assessed in a group situation, using materials and a test booklet that gives simple instructions and a place for students to supply observations. Materials for such tasks are often set out in stations to which students move in a given time interval. The item chosen to illustrate this format comes from the SISS process test for fifth graders (SISS, 1986).

Two small, familiar plastic figures are given to the child to examine. The child must make observations, select relevant attributes from these, and then name the attributes in writing. The reading requirement is limited. The student may ask the test monitor to read a word but *not* to explain it.

Observing Plastic Animals

Two plastic animal specimens are on display before you. Look at them carefully.

1. List three ways in which you can see that they are alike.

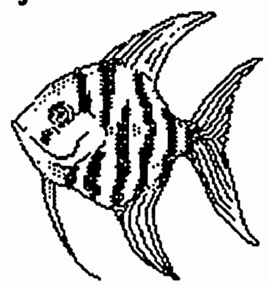
2. List three ways in which you can see that they are different.

Figure 2.12. Observing Plastic Animals

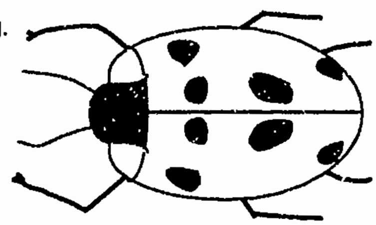
The next examples of items are from a Michigan test for fourth graders (MEAP, 1985). It is a paper and pencil test containing a sampling of content and skills included in the fourth grade curriculum. The items selected for illustration here are taken from the set intended to assess science process skills (Figure 2.13). The specific skill is identified in the heading (Identifying Properties of an Object) but it is not included in the student booklet, where process items are mixed with those designed to assess content. The student does not supply the attributes but instead selects the appropriate ones from a set given. No materials are used; only diagrams of the object. The child needs to be familiar with the object that is represented, to recognize what the diagram represents, and to be able to visualize it. There is also a minimum reading requirement. The number of attributes that a child is required to attend to is limited to those given.

Identifying Properties of an Object

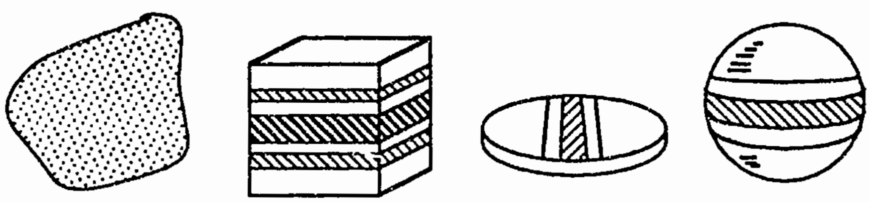
1. Find three properties of this fish.
   A. soft, curvy, spotted
   B. rough, red, hard
   C. soft, slippery, square
   D. soft, curvy, striped



2. Find three properties of this ladybug.
   A. spotted, round, black head
   B. spotted, square, black body
   C. spotted, smooth, has four legs
   D. round, striped, has six legs



3. Find the object that is <u>flat</u>, <u>round</u>, and <u>striped</u>.



A B C D

Figure 2.13. Identifying Properties of an Object.

Classifying

This first example is an assessment item from SAPA (AAAS, 1968). It is intended for somewhat older children (Level C) than the last SAPA illustration and assesses the classification skill. It is similar in form to the other example with teacher demonstrations and scripted teacher questions for the competency item (see Figure 2.14). There is limited student interaction with the materials during this appraisal. Most of what is done with the materials is part of a demonstration by the test administrator. The child needs to be familiar with the concept of change of state of matter; the name of an attribute involved; and, given an example of the attribute, to be able to name it.

Competency Measure for Classifying

(Individual score sheets for each pupil are in the Teacher Drawer.)

Make the following advance preparation: Put 12 to 15 g (about one tablespoon) of citric acid dissolved in 45 ml (about three tablespoons) of warm water in a plastic sandwich bag, and tightly seal off the bottom section of the bag with a rubber band. Then put 5 g (one teaspoon) of baking soda in the open part of the bag, and tightly seal it with a second rubber band. Save this equipment for Tasks 4-8.

TASKS 1, 2 (OBJECTIVES 1, 2): Show the child a test tube in which you have put about 5 ml (one teaspoon) of raw egg white. Say, "Classify the substance in the test tube and give the reason for your classification." Let him handle the test tube if he wishes to. Put one check in the acceptable column for Task 1 if he classifies the substance as a liquid. Put one check in the acceptable column for Task 2 if he states that he can see the flat surface of the substance, or that the substance has no shape of its own and takes the shape of the container, but has a size of its own.

TASK 3 (OBJECTIVE 1): Pour 25 to 30 ml (about two tablespoons) of hot water (71° to 80° C, or 160° to 180° F) along the inside surface of the test tube containing the egg white. Holding he test tube at the top, rotate it between your thumb and forefinger about twenty times. A "strand" of white solid will appear. Ask, "Should the substance or substances in the test tube be reclassified?" Put one check in the acceptable column if the child says, "Yes, now there is a liquid and a solid in the tube."

TASKS 4-6 (OBJECTIVES 1, 2): Hand the child the plastic bag you prepared previously. Ask, "How would you classify the substances you can see in this bag? Also tell me why you make the classifications you do." Put one check in the acceptable column for Task 4 if he says that there is a liquid in one part of the bag and gives the same reason he gave for Task 2. Put one check in the acceptable column for Task 5 if he says the other substance he can see in the bag is a solid. Put one check in the acceptable column for Task 6 if he says the solid substance has a size and shape of its own.

TASKS 7, 8 (OBJECTIVE 1): Cut the rubber band that divides the two portions of the bag and ask the child to shake the bag. Say, "Now classify the substances in the plastic bag." Put one check in the acceptable column for Task 7 and another for Task 8 if he states that there is a liquid and a gas in the bag.

Figure 2.14. Competency Measure for Classifying

The Assessment of Performance Unit (APU) was developed in Britain to assess process skills of a national sample of students (Harlen, Black & Johnson, 1981). The instrument designed for use with eleven-year-olds uses a variety of methods to test classification skills, e.g., using a branching key. This particular skill also requires the skills of observing and using observations. While all are given in a group practical situation, some make use of actual objects, some include photographs or diagrams of objects, and some present a film segment of a phenomenon. Questions are usually of the supply type; the student has to provide the answer, not respond to a pre-selected set. Figure 2.15 assesses the skill of using a branching key with colored drawings being provided.

Using Observations with a Branching Key

Material provided for each pupil: Colored drawings of 3 caterpillars with 'head' labelled on one

Caterpillar Key

1.	Main color on the body is green	If Yes, go to 2
	No large area of green on body	If Yes, go to 4
2.	Yellow slanting stripes along body	If Yes, go to 3
	Stripes on body are blue or white	If Yes, go to 6
3.	Tail and lower half of body is paler in color than the rest	If Yes-Lime Hawk
	Tail and body are the same color all over	If Yes-Poplar Hawk
4.	Large and small yellow spots along the sides of the body	If Yes, go to 5
	Small brown spots along the sides of the body	If Yes-Pine Hawk
5.	Head is red	If Yes, go to 7
	Head is black	If Yes-Striped Hawk
6.	Hook at end of body blue in color...............	If Yes-Eyed Hawk
	Hook at end of body red in color	If Yes-Silver Striped Hawk
7.	Red stripes going from head to tail	If Yes-Spurge Hawk
	No stripes along body	If Yes-Bedstraw Hawk

Question Page

Use the statements in the CATERPILLAR KEY to find the names of the three caterpillars A, B, and C.

Start with caterpillar A and the statements at number 1. Find which statement fits caterpillar A. When the answer is YES you will find either the name of the caterpillar or the number of the next set of statements to go to.

Write the name of each caterpillar and the numbers of the statements you used to find the name. (Number is written for you).

a) Caterpillar A is: ______________

Statements used: 1, ______________

b) Caterpillar B is: ______________

Statements used: 1, ______________

c) Caterpillar C is: ______________

Statements used: 1, ______________

Figure 2.15. Using Observations with a Branching Key.

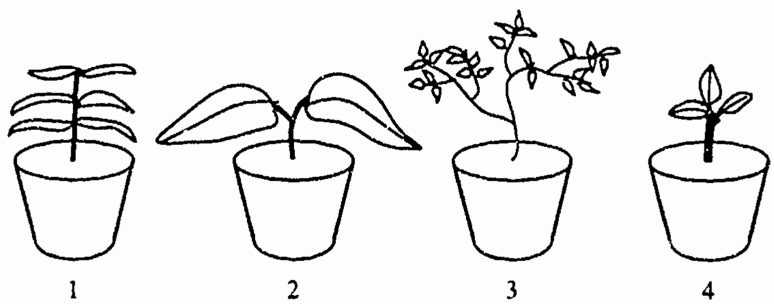
There is a large number of pencil and paper items designed to assess a wide range of classification skills. Two are given in Figure 2.16. They are both from the same test (Jefferson County, 1982). The first requires the child to know the property in question and to be able to identify the one material that does not have that property. The second item asks the student to serrate the leaves in size in the direction specified.

Evaluating Classification Skills

1) All of the following will float on water EXCEPT :

A. cork.
B. oil.
C. wood.
D. marble.

2) Arrange the plants in order from the <u>longest</u> leaves to the <u>shortest</u> leaves.



A. 2, 1, 4, 3
B. 3, 1, 4, 2
C. 1, 4, 2, 3
D. 2, 3, 1, 4

Figure 2.16. Evaluating Classification Skills

Using Symbols

There are numerous ways in which symbols are used in science as well as in other domains and a great number of subsidiary skills are involved. Recording and communicating data require skills such as reading and expressing data in charts and graphs in an organized manner appropriate to the particular task. It takes no great leap of the imagination to see how items could be written using the varied formats as described in the above categories: from observation of performance with materials to exclusively paper and pencil items that are multiple choice. Two sets of illustrations have been chosen for the category of using symbols. The first set of multiple choice items are from the SISS achievement test (SISS, 1984). They involve reading a chart, matching one of the temperatures on the chart to a diagram of a thermometer (which actually is just a number line), and making a simple inference based on the information presented in the chart (Figure 2.17).

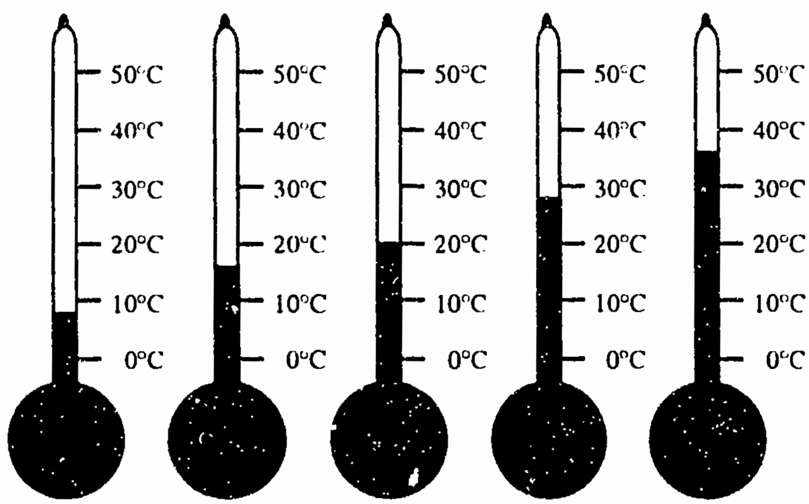
Evaluating Table and Interpretation Skills

The next three questions refer to the following table which shows some temperature readings made at different times on three days.

	6 a.m.	9 a.m.	12 noon	3 p.m.	6 p.m.
Monday	15°C	17°C	20°C	21°C	19°C
Tuesday	15°C	15°C	15°C	10°C	9°C
Wednesday	8°C	10°C	14°C	14°C	13°C

1) When was the highest temperature recorded?
   A. Noon on Monday
   B. 3 p.m. on Monday
   C. Noon on Tuesday
   D. Noon on Wednesday
   E. 6 p.m. on Wednesday

2) Which of the following shows the temperature at 6 a.m. on Wednesday?



A B C D E

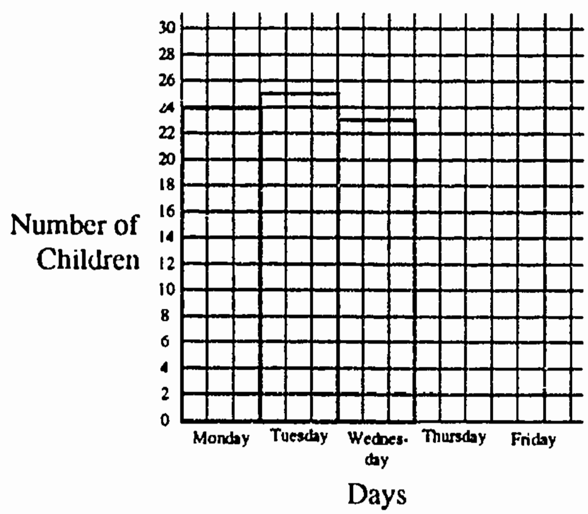
3) On one day a cool wind began to blow. When do you think this happened?
   A. Monday morning
   B. Monday afternoon
   C. Tuesday morning
   D. Tuesday afternoon
   E. Wednesday afternoon

Figure 2.17. Evaluating Table and Interpretation Skills

The next illustration consists of three interesting items on graphing which were on one of the APU tests for eleven-year-olds (Harlen, Black & Johnson, 1981). They vary in how much of the graph has already been done and how much the student must supply. For "A" the child needs only to enter data on the graph; "B" gives the child less help in constructing the graph; and "C" the least help. Item "C" presents only the data to be used and a grid for the graph. The items range in difficulty with "A" being the easiest and "C" the most difficult, requiring the mastery of more of the skills of graphing.

Evaluating Graphing Skills — Minimal Difficulty

A. This graph shows the number of children staying to dinner on the first three days of one week.



a) On Thursday 20 children stayed to dinner. Add this to the graph.
b) On Friday 26 children stayed to dinner. Add this to the graph.

Figure 2.18A. Evaluating Graphing Skills—Minimal Difficulty

Evaluating Graphing Skills — Moderate Difficulty

B. This table shows how far each of these children can swim:

Name	Number of Lengths
Mike	2
Dennis	Can't swim
Judy	2
John	1
Mary	1
Jill	2
Ian	1
Sue	1
Jane	Can't swim
Alan	1

Draw a bar chart to show how many children can swim each distance.

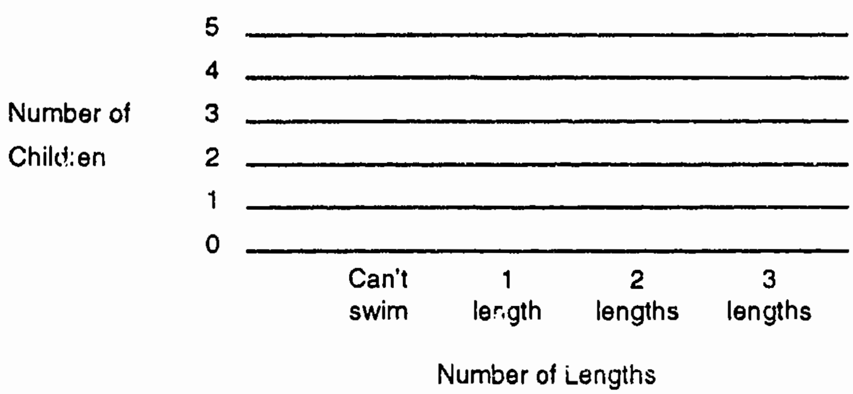


Figure 2.18B. Evaluating Graphing Skills—Moderate Difficulty.

Evaluating Graphing Skills — Greater Difficulty

C. Richard measured his bean plant every week so that he could see how fast it was growing. He started (0 weeks) when it was just 5 cm high. These were the heights for the first four weeks:

0 weeks -----------	5 cm
1 week -----------	15 cm
2 weeks -----------	30 cm
3 weeks -----------	40 cm
4 weeks -----------	45 cm

Draw a graph to show how the height changed with time.

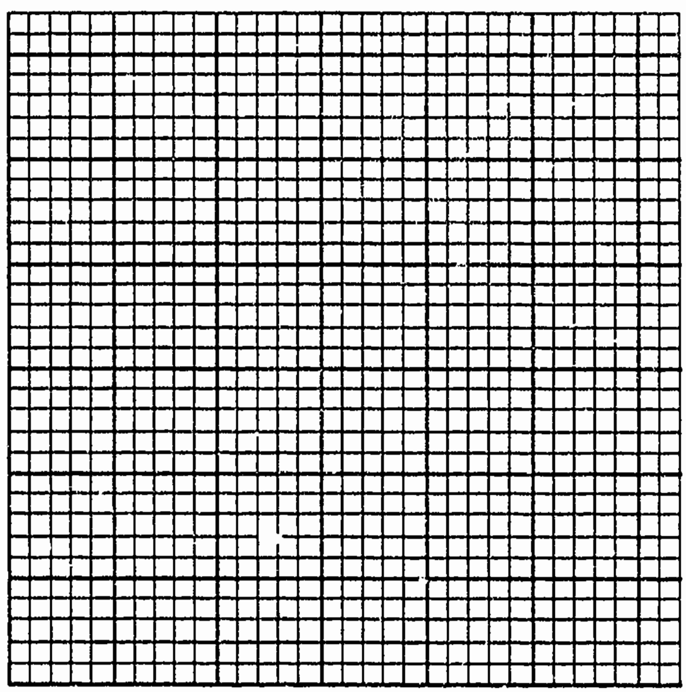


Figure 2.18C. Evaluating Graphing Skills—Greater Difficulty

Predicting

The last example illustrates a task that has been used with second grade students (Chiarello, 1989). First, the students predict for each of the six situations whether it would balance or not. After completing that phase, they are asked to manipulate the blocks to see if their predictions were accurate.

Evaluating the Skill of Predicting

Predict which ones would balance. After you have finished your predicting, do the test.

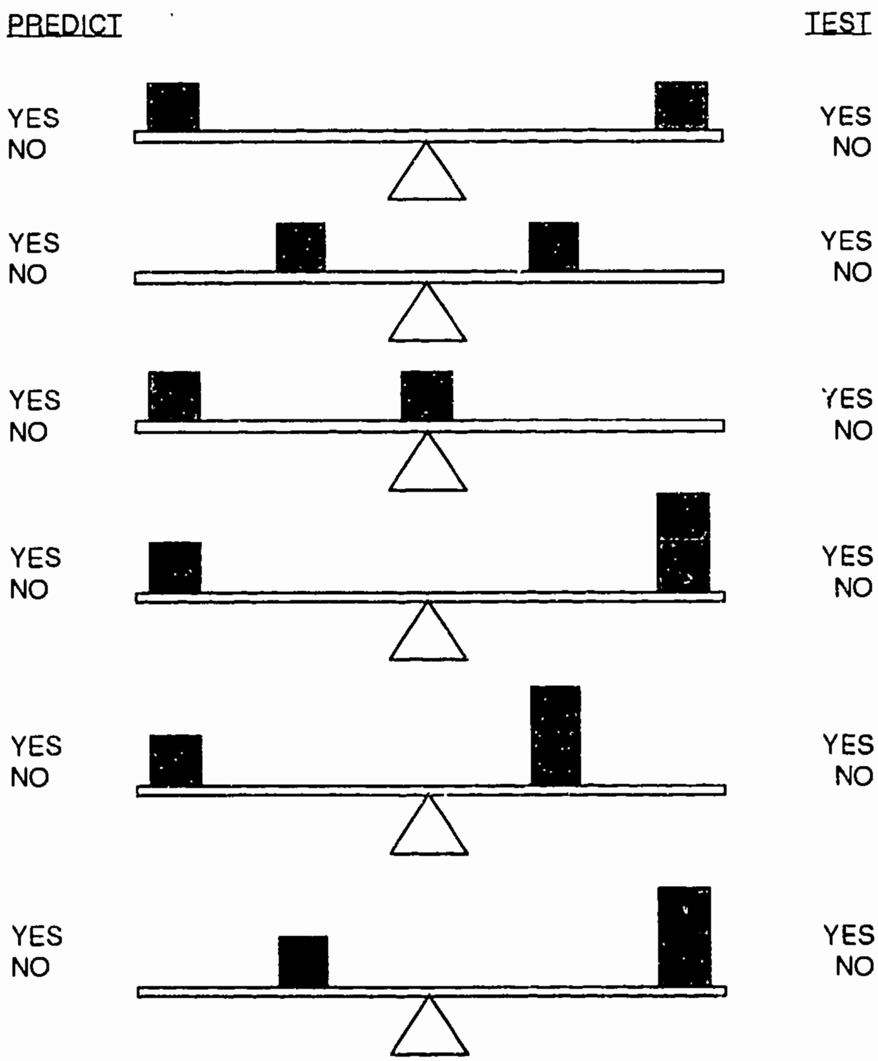


Figure 2.19. Evaluating the Skill of Predicting.

A sample of assessment items for skills that are usually considered part of the common curriculum have been illustrated in this chapter. They are often assessed separately, although they could also be assessed in the context of an investigation. Other skills more specific to problem-solving or investigating will be discussed in Chapter 4, "Problem Solving."

CHAPTER 3

ASSESSING CONCEPTS

Introduction

Most science educators cite processes and concepts as the major outcome of school science programs at all levels. Some people have expressed the view that elementary school science should be oriented only around process outcomes, not content. When pressed, most of these people could be described as being "anti-recall." For example, they agree that students should know about cells, forces, rock cycles, and changes of state. Some of the programs and experimental projects have stressed one set of outcomes over the other, but most recognize the need to include both the procedures (processes) and the products (concepts) of science.

As scientists and young people observe and describe natural phenomena, they find it necessary to group and relate observations and facts to form a simpler view of nature. Through this grouping and relating activity they are able to develop an integrated picture from initially separate facts. This is the process of concept building. Concepts help simplify the understanding of past, present, and future experiences because individual facts become a piece of a mosaic. Individuals form concepts based on their specific sets of experiences and grouping procedures. With the rapid growth of knowledge in the fields of science, such summarizing or generalizing of ideas is absolutely essential. Because of their comprehensive nature, concepts enable the possessor to have some grasp of a much larger field than he has personally experienced. They also facilitate the interpretation and assimilation of new information and observations.

Concepts and Teaching

The National Science Teachers Association has been instrumental in highlighting the importance of concepts in elementary school science. The "big ideas" or major concepts of science have been the organizational structures for many curriculum projects and guidelines. One example is the Science Curriculum Improvement Study (SCIS), which is organized into a series of major conceptual themes for the life sciences and the physical sciences. SCIS also addresses process skills and attitudes in addition to the major science concepts, which are considered the organizational structures of the program. The New York State Elementary Science syllabus (NYSED, 1985) is organized around a similar scheme of interrelating the concepts of the life and physical sciences (see Figure 3.1).

Concept Scheme for New York Elementary Science Syllabus

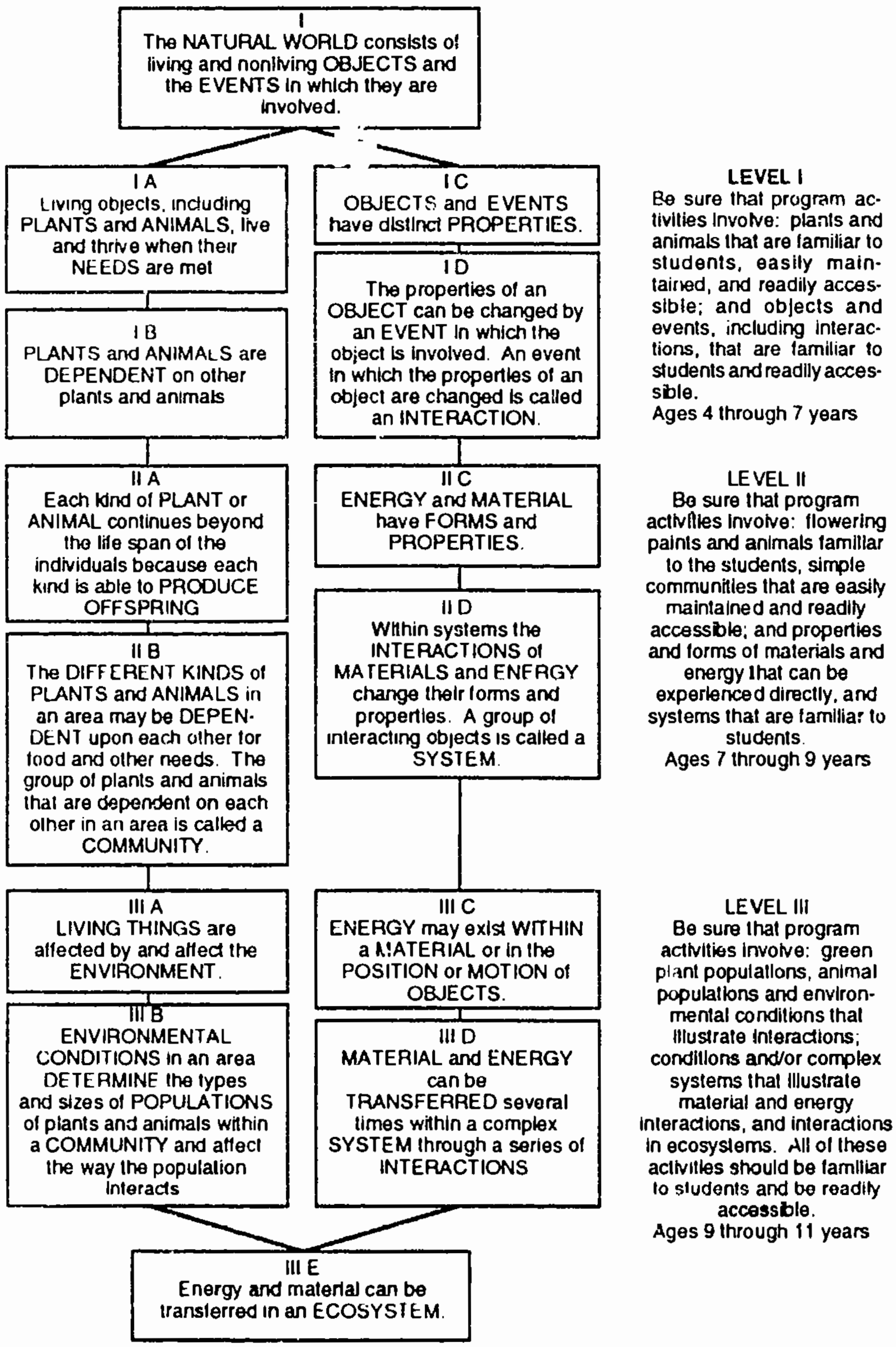


Figure 3.1 Concept for New York Elementary Science Syllabus

Most elementary science instructional programs also are organized around major science concepts or themes. The Addison Wesley program, STEM, is built around the concepts of space, time, energy, and matter. The State of Wisconsin recommended that science programs be organized around the concepts of diversity, change, continuity, interaction, organization, and limitation. Instead of a single dimension or list, the Wisconsin guide created a three-dimensional matrix of science outcomes. The other two dimensions are the science areas (biological, physical, and earth sciences), and the science processes (observing, classifying, inferring, formulating models, etc.).

Most of the work with conceptual schemes in science has been oriented toward developing curricula and preparing instructional materials, rather than evaluating students or programs. In terms of the assessment of elementary science outcomes, most efforts revert to the traditional science divisions (life and physical) and to detailed subdivisions within these categories. Some efforts have added an earth science area, as well as integrated topics. The elements within these schemes become the labels for the familiar units or topics on which so many elementary science programs are based. These topics include concepts such as leaves, balloons, dinosaurs, rocks, water, machines, and electricity.

Model for Assessing Concepts

It is clear that concepts are more than a collection of information or facts. Most people agree that a concept exists when two or more objects or events are grouped together on the basis of some common feature or property. These common properties are called the relevant attributes of the concept. Concepts commonly become labelled by a word or phrase which then represents the concept. A definition of a concept is an important means for characterizing or communicating a concept to others.

Concepts vary in terms of their level of abstraction and their potential use. Three categories of concepts have been cited by some writers. *Classificatory* concepts (e.g.. cloud) are largely useful for describing phenomena, while *correlational* concepts (e.g., structure and function of organs) facilitate prediction. Similarly, *theoretical*, or abstract, concepts (e.g., molecule) are used to explain observations and phenomena.

Researchers at the Wisconsin Research and Development Center for Cognitive Learning developed a model (Frayer, Frederick & Klausmeir, 1969) for testing levels of concept mastery. The model consists of the following 12 tasks:

1. Given the name of an attribute, select an example of the attribute.

2. Given an example of an attribute, select the name of the attribute.

3. Given the name of a concept, select the example of the concept.

4. Given the name of a concept, select the non-example of the concept.
5. Given an example of the concept, select the name of the concept.
6. Given the name of the concept, select the relevant attribute.
7. Given the name of a concept, select the irrelevant attribute.
8. Given the meaning of a concept, select the name of the concept.
9. Given the name of a concept, select the meaning of the concept.
10. Given the name of a concept, select the supraordinate concept.
11. Given the name of a concept, select the subordinate concept.
12. Given two concepts, select the principle relating them.

Classificatory Concepts

The above scheme was used to develop test items for classificatory concepts: those based on a classification or description of objects and phenomena in nature. A project (Voelker & Sorenson, 1971) to construct test items on classificatory science concepts appropriate for intermediate grade children selected the following concepts:

Biological	Earth	Physical
Bird	Cloud	Conductor
Cell	Core (Earth)	Evaporation
Fish	Fossil	Expansion
Heart (Human)	Glacier	Friction
Invertebrate	Meteor	Liquid
Lens (Eye)	Moon	Melting
Lungs	Planet	Molecule
Mammal	Sedimentary Rock	Solid
Muscle	Volcano	Sound
Pore (Skin)	Wind	Thermometer

Their scheme for testing levels of concept mastery is best illustrated by a "concept analysis" and a sample of test items for one concept. The concept chosen is "evaporation."

An Example of a Concept Analysis for Evaporation

Definition — Evaporation is the process by which a liquid changes to a gas as particles escape from the surface of the liquid.

Supraordinate Concepts(s). process

Coordinate Concept(s): condensation, melting, freezing

Subordinate Concept(s): slow evaporation, rapid evaporation

Critical attributes that differentiate the target concept from the supraordinate concept (or coordinate concepts if a supraordinate has not been identified:

In evaporation:
1. a liquid changes to a gas
2. particles escape from the surface of the liquid

Other attributes that are relevant but not criterial for the target concept (the attributes of the supraordinate need not be specified):

Other attributes relevant to evaporation are those of its supraordinate, process, such as the process:
1. involving molecular motion
2. involving energy

Irrelevant attributes of the target concept (attributes which vary among instances of the target concept) include the following:

Irrelevant attributes of evaporation include:
1. speed of occurrence (e.g., slow, rapid)
2. color of the liquid

Relationship with at least one other concept (Preferably this relationship should be a principle. It should definitely not be a direct supraordinate/subordinate relationship, a relationship involving a criterial attribute, or a relationship involving an example.):

Evaporation can cause *cooling.*

Concept examples include the following:

evaporation caused by sun shining on the earth
steam escaping from a pan

Concept non-examples include the following:

condensation of water on the outside of a glass
snow storm
rain storm

Some examples of items assessing levels of concept mastery

Level 5 This picture shows:
- A. burning.
- B. melting.
- C. condensation.
- D. evaporation.

Level 8 The process by which a liquid changes to a gas as particles escape from the surface of the liquid is called:
- A. burning.
- B. condensation.
- C. evaporation.
- D. melting.

Level 9 Evaporation is a process by which:
- A. a substance changes in volume because its particles move farther apart.
- B. a solid changes to a liquid because of increased motion of the particles
- C. a liquid changes to a gas as particles escape from the surface of the liquid

Level 12 What is true about evaporation and cooling?
- A. Cooling is necessary for evaporation to take place
- B. Cooling speeds up evaporation.
- C. Evaporation can cause cooling.

Figure 3.2. An Example of a Concept Analysis for Evaporation

Correlational Concepts

Correlational concepts are those that invoke a relationship between two variables. The above model can be used here as well.

Some examples of correlational concepts are:

- The parts of cells have specific functions.
- Plant and animal cells differ in structure, depending on their function.
- The activities associated with life are carried on in the cell.
- Organisms differ in size, depending on the number of cells possessed.

Theoretical Concepts

Theoretical concepts are ideas created by humans to explain phenomena but are not based on direct human experiences; they are abstractions. Examples of theoretical concepts are the following:

- DNA is the important molecule concerned with regulation of cellular activity.
- The particles which make up matter are in motion.
- Matter is made up of particles.
- The particles which make up matter have spaces between them.

Concepts and Children's Experiences

It is obvious that concepts exist along a concrete/abstract continuum. For example, the classificatory concepts are generally based on observations made directly with one of the five senses. At the other end of the spectrum, theoretical (abstract) concepts are based on models or theories that can not be directly perceived through the senses.

Concepts can also be analyzed along a familiar/not familiar dimension. Familiar concepts are those based on experiences and information common to youngsters at a particular level of schooling in a particular setting. For instance, third grade youngsters in the northeastern United States should be familiar with concepts related to grass, trees, dogs, cats, the sun, the moon, clouds, rain, and snow. On the other hand, they might *not* be familiar with giraffes, llamas, kangaroos, magnets, molecules, acceleration, density, reproduction, and genetics. Their lack of familiarity may be based on the item's complexity, its abstractness, or the lack of first-hand experiences.

Concepts and Cognitive Development

A major concern with testing the concepts held by youngsters in the elementary school is their level of cognitive development. Piaget has described children's thinking along a continuum from sensorimotor (birth to two years), preoperational (two years to seven years), concrete operational (seven years to eleven years), and formal operations (eleven years and beyond). These ages are based on results from considerable research, but variations do exist for individuals in a given culture and for groups of individuals in different cultures. The following illustration (Figure 3.3) shows the characteristics of youngsters at different Piagetian stages for the float/sink (density) concept (Shayer & Adey, 1981).

Topic		
Floating and Sinking (Density)	**2A early concrete** At this level, Mass, Weight, Volume, and Density are still "collapsed" in a global notion of "heaviness"; knows that wood will float, iron will sink, but without a general explanation available, he can only learn a series of individual facts about materials.	**2B late concrete** Specific theories of floating will be tested, and weight differentiated from mass as a variable. Volume will only partly be conceptualized, and so the weight/volume relationship will not yet be used as an explanatory tool. Different "heaviness" of materials will be differentiated from "bigness." "A small or a large piece of plasticine will both sink, because the stuff is the same, with the same heaviness."
	3A early formal Volume conceptualized and displacement seen to be a function of volume, not weight. Weight/volume relationship will be utilized to generate hypotheses in the floating/sinking problem. Complete solution, including density of liquid, unlikely to be discovered, but rules about relative density can be learned. "You can find out if two things are the same substance by seeing if their weight/volume ratio is the same."	**3B late formal** Can handle relationship between, say, density, mass, and spacing of particles. Could formulate a theory of floating, relating density of solid to density of liquid, or is likely to find that the clue to the floating and sinking problem is the weight of displaced liquid.

Figure 3.3. The Concept of Density Classified by Developmental Levels

The example spans the range of likely stages experienced with elementary-school-aged children, from early concrete to late formal. In assessing youngsters, we must check to be sure that we are focusing on thinking that is reasonable to expect for youngsters of a particular age or grade level if we are assessing achievement rather than cognitive development. As one can see from the illustration, thinking that requires relationships and several variables impinging on a phenomena simultaneously will only be possible with children at the "formal" level of development.

These levels of cognitive functioning become very useful for establishing "benchmarks" for student understanding with respect to specific concepts as was used earlier with science skills. The following illustration (Figure 3.4) on the concepts of force and energy shows expected explanations for children, with the reasoning of less sophisticated children on the left and on the right the fuller explanation of the youngsters functioning on the abstract (formal) level (Harlen, Darwin & Murphy, 1978).

Levels of Mastery — Concept of Force

Generally tries to explain the movement of an object in terms of its own will or ability to move rather than the action of forces on it.	Explains the movement of an object in terms of one or more forces acting on it, but does not identify forces acting when no movement takes place.	Can identify forces acting on an object whether in motion or remaining still. Appreciates that forces acting are equal and opposite when an object is not moving.

Levels of Mastery — Concept of Energy

Considers energy as if it were a substance which is created and lost somewhat magically, without any continuity between one form and another.	Identifies energy in various forms and recognizes its conversion from one form to another but considers that it can be created at some point in a chain and used up at others.	Describes the changes of form of energy which take place in simple energy chains, recognizing that when it seems to disappear in one form it reappears in some other form.

Figure 3.4. Levels of Mastery—Energy

A Cognitive Taxonomy

Few ideas have had as great an impact on education as the Bloom taxonomy of cognitive objectives (Bloom, et. al., 1956). The six levels of the Bloom taxonomy, knowledge, comprehension, application, analysis, synthesis, and evaluation, have been used widely in curricular and assessment activities. The authors believe that a "collapsed" taxonomy consisting of fewer categories is more defensible and simpler to use. For the purpose of this monograph, we have chosen the categories of knowing, using, and extending. The "using" category generally incorporates the middle Bloom categories of comprehension and application, while the "extending" category consists of the upper three categories of analysis, synthesis, and evaluation.

The taxonomic (or ladder) aspect of these categories is portrayed as a triangle, as illustrated below.

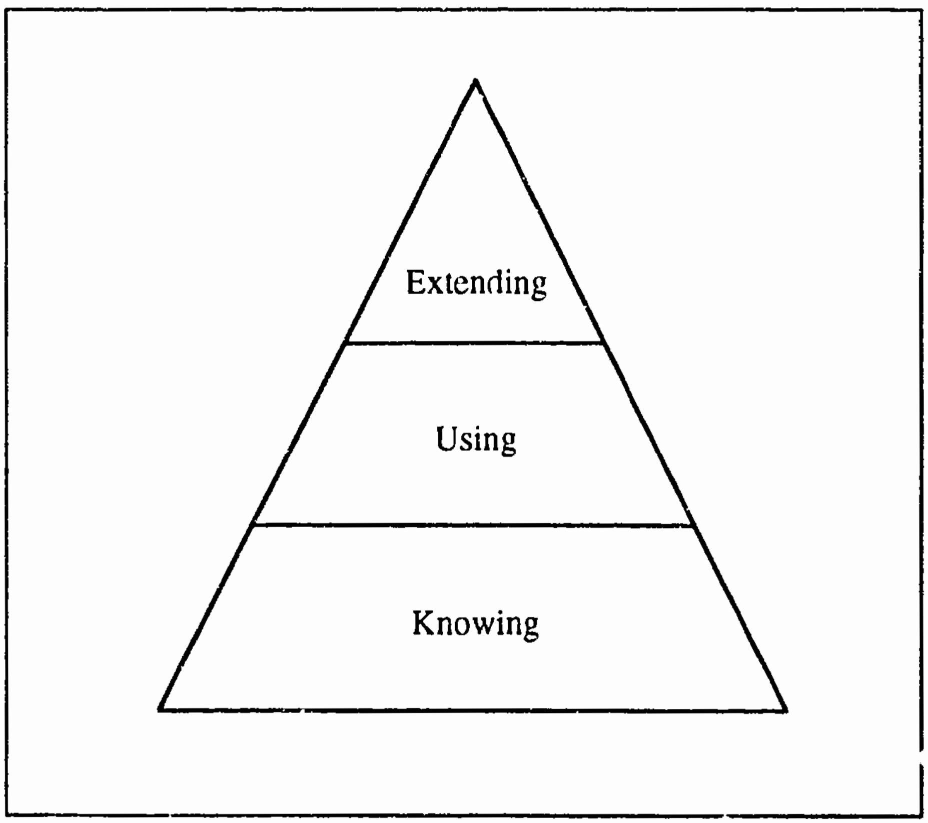


Figure 3.5. Modified Bloom Cognitive Taxonomy.

The configuration of the layers implies that using objectives are generally based on a foundation of appropriate knowing objectives. Similarly, before a person can effectively extend ideas, a base of knowing and using must be mastered.

Educators have cautioned against the over-simplification of taxonomies of educational objectives. One must be sensitive to the dependence of these levels on the instruction experienced by the students who are being assessed. If the instructional program presents a specific "extending" situation in great detail, then to use that identical situation in the assessment mode creates problems. It is quite likely, in that case, that the resulting test item should be categorized as a "knowing" type. Extending this argument further, it is clear that not all students receive the same instructional experiences. Therefore, one must exercise great caution "labeling" an item as generally being at a particular level, because this is based on an assumption of a set of instructional experiences that are "universal."

Test Blueprint

A widely used tool for developing tests which assess a unit or a course in a balanced or fair manner is the test grid or blueprint. This grid is usually composed of two dimensions: one describing the *content* area and the other the appropriate *levels of objectives* for these students. In addition to the listing of the content and objective categories, the test grid is based on a determination of the relative emphasis of each sub-category within these dimensions. The example below (Figure 3.6) is based on an elementary science program organized by the life, earth, and physical science areas. For the purpose of illustration, we assumed that the physical and the earth science areas are of equal importance, so each has an "emphasis rating" of 25%, while the life science area is of more importance with a 50% rating. That means that within each assessment of this program one-fourth of the testing questions should be on physical science content, one-fourth on earth science, and one-half on life science information. When one is producing a test grid for a specific program, the emphasis ratings should be based on a measure of the time spent on each area. One could simply count the number of days or periods spent on each science area and divide by the total amount of time spent on all science subjects.

The emphasis of the separate levels on the objective dimension (knowing, using, and extending) is not quite as easy to accomplish as is the content emphasis. An individual teacher or a group of teachers could arrive at an estimate of the emphasis of the different levels by examining lesson plans, text books, and curriculum guides. Even though this rating is more of an "estimate," it is still an important parameter for the planning of the test. After this rating, or "desired state," has been determined for several consecutive years, it becomes "fine-tuned" to the classroom experiences of the students. In this hypothetical example the "knowing" and "using" outcomes are equally stressed, with each accounting for 40%, while the "extending" outcomes are somewhat less emphasized, as indicated by the 20% rating.

		CONTENT		
		Life	Earth	Physical
OBJECTIVE		50%	25%	25%
Extending	20%	10%	5%	5%
Using	40%	20%	10%	10%
Knowing	40%	20%	10%	10%

Figure 3.6. Example of a Test Blueprint.

Once the emphases have been determined for both the content and objectives dimensions, one can calculate the value for each "box" in the matrix by cross-multiplying the row and the column values. For example, the value for "knowing life sciences" is 40% x 50% = 20%. That is interpreted to mean that 20% of the items within a "balanced" test of this hypothetical science program should test students' ability to "know information within the life science area." The numbers in the boxes in Figure 3.6 represent the percentage of the total test that assess each particular combination of content and objectives.

If a test were to be composed of twenty items, the following distribution of items (Figure 3.7) would be recommended to fit the guidelines of the above example. Note that five "physical items" represent 25% of this twenty-item test. Similarly, the eight "using items" comprise 40% of this test, as recommended by the test grid.

	Life	Earth	Physical	
Extending	2	1	1	4 (20%)
Using	4	2	2	8 (40%)
Knowing	4	2	2	8 (40%)
	10 (50%)	5 (25%)	5 (25%)	20 (100%)

Figure 3.7. Distribution of Items in 20-Item Test.

Illustrative Test Items

The remaining sections of this chapter will consist of sample items with brief comments and descriptions. The first section will include illustrations of test items which make minimal reading demands. While these might be most useful for students in the younger grades, they might also be helpful for youngsters whose first language is not English. The last three sections are organized to include items assessing concepts from the life, earth, and physical science areas. Within these areas, items are grouped by the level of student cognition required: knowing, using, and extending. Readers should recall the cautions stated earlier about the difficulties of categorizing items by objective level. The visuals for the illustrating items are not precisely those of the original materials. They have been modified for storage in microcomputer files and for standardization across examples. The full title of the source of each item is listed at the end of the chapter. In order to simplify the referencing, the following coding system was developed:

Code		Test Source
ISS	=	Second International Science Study
JEF	=	Jefferson County Science Test
MAT	=	Metropolitan Achievement Test
STE	=	Sequential Test of Academic Progress, Level E
STG	=	Sequential Test of Academic Progress, Level G
MAP	=	Michigan Education Assessment Program
O&F	=	Osborne and Freyberg
SCI	=	Science Curriculum Improvement Study
SAT	=	Stanford Achievement Test, Primary Level II
CTB	=	Comprehensive Test of Basic Skills, Level C, Form S
IOX	=	Instructional Objectives Exchange

Items with Minimal Reading Demand

The format used for most test items in this section has the student choose from several pictures the correct response to a question which is read aloud by the test administrator. The first set of items was selected from Level C of the *Comprehensive Test of Basic Skills* (CTB/McGraw Hill, 1973). The corresponding part to be spoken by the person administering the test is listed for each item. The items are illustrated in Figure 3.8 and the oral directions in Figure 3.9.

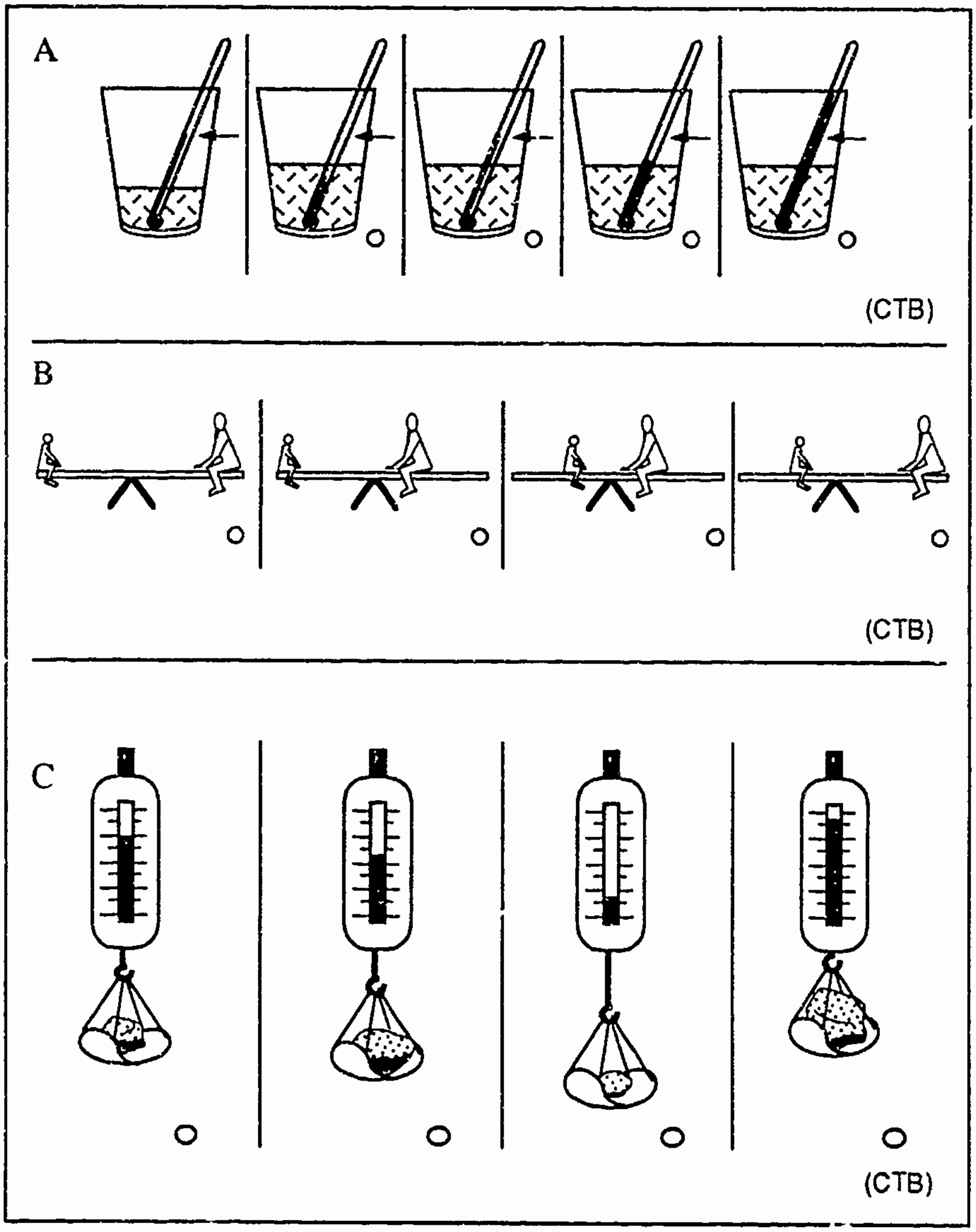


Figure 3.8. Example Test Items with Minimal Reading Demand.

ITEM A
SAY: Item A. Look at the first picture of the thermometer. It is in a glass of cold water. The arrow shows how high the dark line is. Listen carefully to what I say. Heat will make the dark line move up. Some children put some hot water in the glass. Now look at the rest of the pictures in the row. Find the picture that shows what happened to the line on the thermometer when the hot water was added. Mark the circle under the picture.

ITEM B
SAY: Look at Item B. See the boys on the seesaw. One boy is much bigger than the other. Find the picture that shows how the boys have to sit so that the seesaw will balance. Mark the circle under the picture.

ITEM C
SAY: Item C. Listen to what I say. Betty collected four rocks. She put each rock on a spring scale. She knew that the heaviest rock would make the marker on the scale go down the farthest. Find the picture that shows the heaviest rock. Mark the circle under the picture.

Figure 3.9. Verbal Directions for Test Items with Minimal Reading Demand.

These items can be constructed with a wide variety of content and at several levels of cognition - as long as a visual representation or model is possible. Some of these items could assess knowledge of concepts. Others are dependent upon the previous performance of some tasks or activities, such as using thermometers or a spring scale. The "see-saw" item could perhaps be answered based on playground experience or on quite sophisticated proportional thinking.

The second set of items which have visual choices and orally spoken directions are from the *Stanford Achievement Test*, Primary Level II battery (Harcourt, Brace & Jovanovich, 1973). According to the administrative manual, this test is designed for use primarily from the middle of Grade 2 to the middle of Grade 3, but it might be used above or below this level with classes at different ability levels.

As you can see from the printed directions, which are to be read, the administrator pauses between each question, providing time for student thinking and response. That interval of time should be monitored for appropriateness, depending on the motivation of the youngster and the cognitive level of the question.

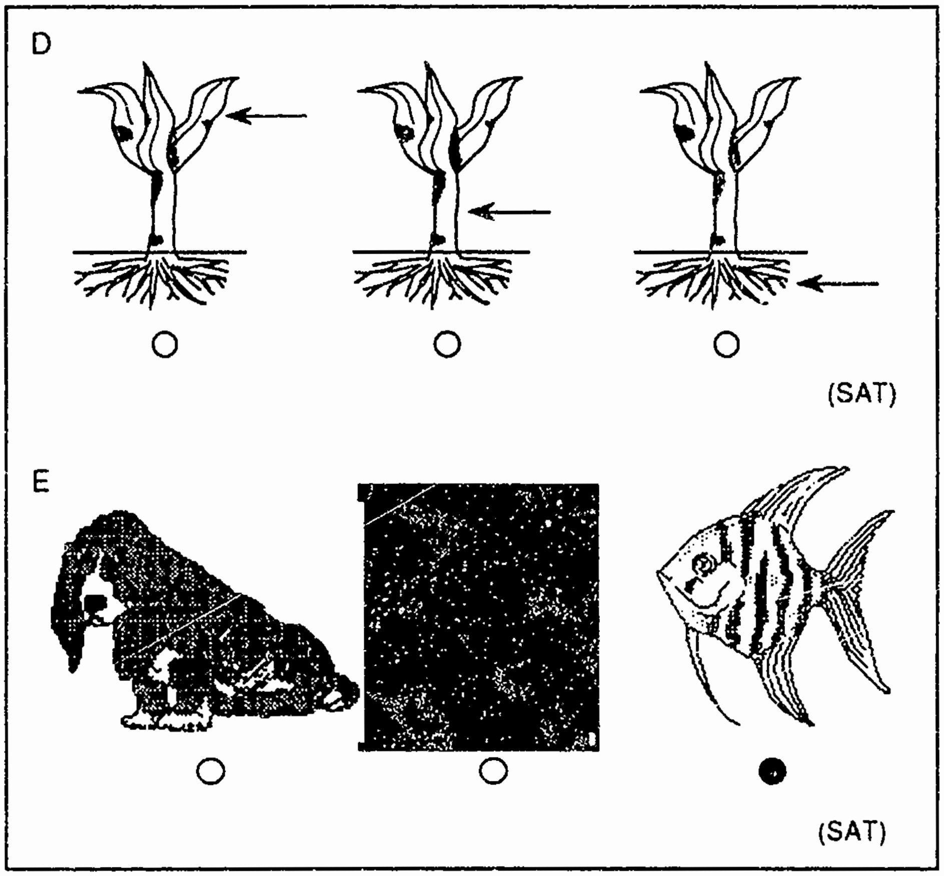


Figure 3.10. Examples of Pictorial Test Items.

Read each question once. Pause about ten seconds between questions. Be sure to read the item number before each question.

D Here are three pictures of a young plant. In which one of these pictures does the arrow point to the stem?

E Which one of these animals does not hatch from an egg outside the parent's body?

Figure 3.11. Verbal Directions for Pictorial Test Items.

Item D assesses the students' ability to "know" what part of a plant is the stem. To correctly answer item E, students must "know" (from class, actual experiences, or books) which of these animals is not hatched from an egg outside the parent's body. It is also conceivable that they could know that two are "egg-layers" and choose the third animal.

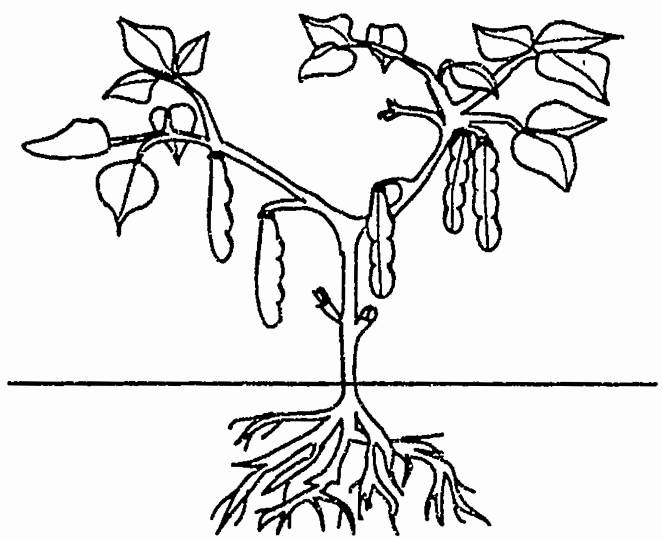
Figure 3.12 with the drawing of a bean plant is from an IOX publication (IOX, 1972). The student has to read a brief description and then two sentences of directions, which require marking a part of the illustration with an "x" and circling another part.

Sample Item:

Directions: Here is a bean plant.

a. Put an "X" on the part of the plant that makes food for the plant.

b. Circle the part of the plant where seeds are made.



Answer:

a. An "X" should be placed on a green leaf.

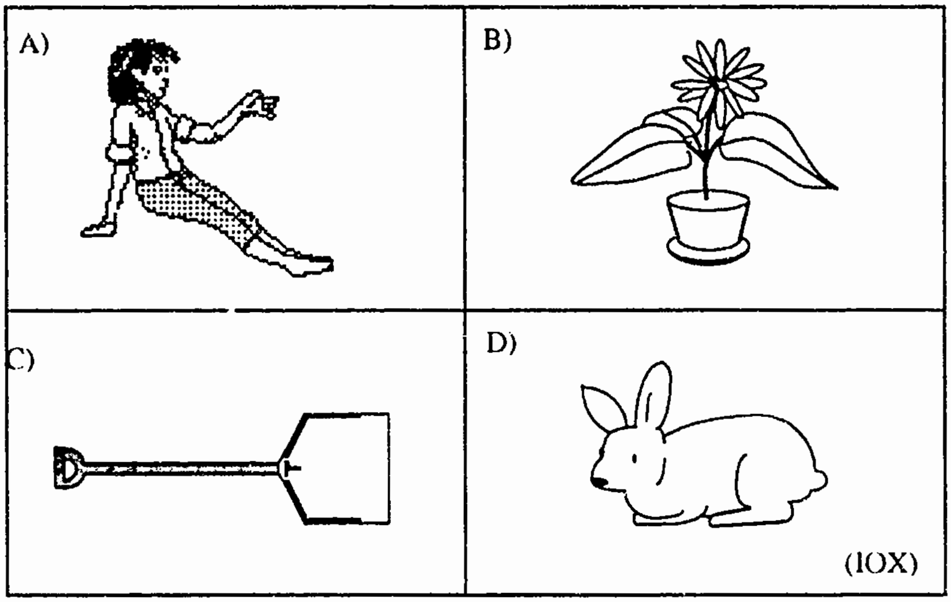
b. A bean pod should be circled.

(IOX)

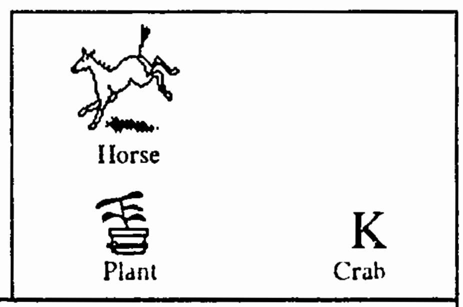
Figure 3.12. Evaluation of a Concept Using a Diagram.

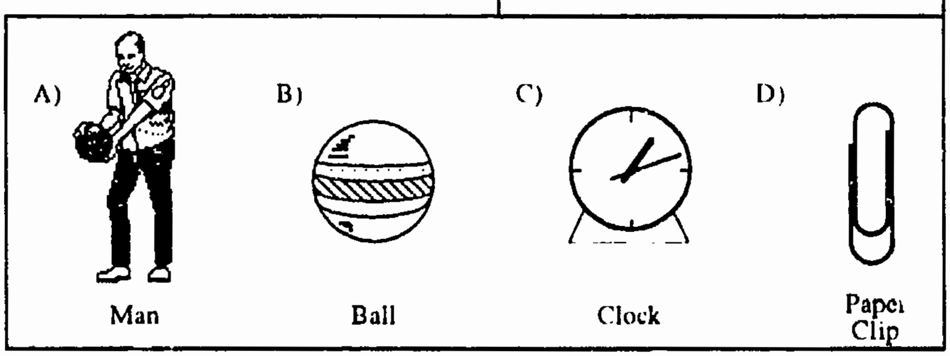
The two life science items of Figure 3.13 are examples of test items with some reading demands, likely understood by most second graders. Knowledge of one key element of all living things, the properties of growth, is assessed in the first item. The second requires the student to determine that the horse, plant, and fish are all living things and then to select from the choices another living thing, man.

1. Which does NOT grow?



2. Which of these fits best with the horse, crab, and plant?





(IOX)

Figure 3.13. Evaluation of Concepts Using Diagrams.

Earth Science Items

Knowing Earth Science

The items displayed in Figure 3.14 were selected to illustrate the assessment of earth science topics at the knowing level. In each item a single word or very short phrase typifies the answer.. These items measure definition of a term (star, earthquake), the largest of a group of bodies in our solar system, and the name of the object which the Earth orbits. This information must be recalled or recognized by the student to answer each question correctly.

1) What is the Sun?

   a) a planet
   b) a star
   c) a comet
   d) a moon
   e) a meteor

   (ISS)

2) The release of pressure along a fault or crack in the Earth's crust results in:

   a) an eruption
   b) a volcano
   c) an atomic reaction
   d) an earthquake
   e) a tornado

   (ISS)

3) Which of the following is the largest body?

   a) Earth
   b) Mars
   c) the Moon
   d) the Sun
   e) Venus

   (JEF)

4) Each year, the Earth move once around :

   a) Mars
   b) Venus
   c) the Sun
   d) the Moon
   e) all of the other planets

   (JEF)

Figure 3.14 Examples of Earth Science Items at the Knowing Level

Using Earth Science

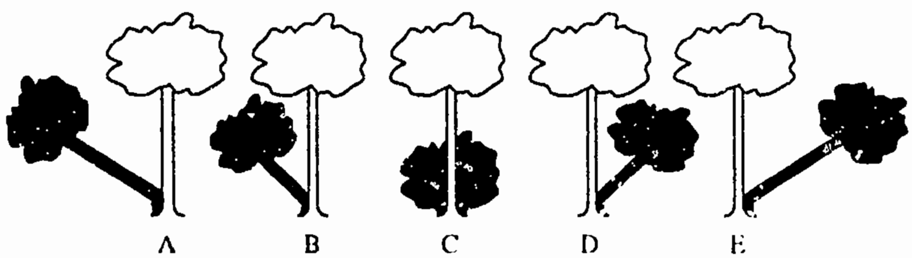
The using of earth science topics is illustrated by items in Figure 3.15. A higher level of cognitive functioning is required to respond to these questions. These items could have been taught directly, and then these items would have assessed at the knowing level. In the first item students must deduce that the moon is "seen" by light reflected from the Sun, since the moon does not generate light

In the next item the tree's shadow at noon can be determined by recalling the direction of shadows of self or objects at noon and comparing that to the tree examples. The compound item involving comparisons of the Earth and moon requires students to determine separately whether each statement is true for the Earth, for the moon, for both, or for neither. The rather sophisticated pattern response makes substantial logical demands on elementary school youngsters. In the fourth example the seemingly simple observation that day follows night is explained by the continual rotation of the Earth, exposing different sections of the Earth to the Sun's light when "facing" the Sun and then causing each section to be dark when the Earth faces "away" from the Sun. The last item, related to life on the Moon, requires students to apply human needs to a "non-Earth" environment.

1) The Sun is the only body in our solar system that gives off large amounts of light and heat. Why can we see the Moon?

A. It is reflecting light from the Sun
B It is without an atmosphere
C. It is a star.
D It is the biggest object in the solar system.
E It is nearer the Earth than the Sun (ISS)

2) At different times during a sunny day a tree was seen to have cast shadows of different lengths as shown in the diagrams below. Which diagram shows the shadow at mid-day (12 noon)?



A. diagram A D. diagram D
B. diagram B E. diagram E
C. diagram C (ISS) (ISS)

3) Here are some sentences about the Earth and Moon that may or may not be true. Read each sentence and decide whether it is—

E true of the Earth only G true of both the Earth and Moon
F true of the Moon only H true of neither the Earth nor Moon

It has oceans of water on it
E F G H
It is larger than the Sun.
E F G H
It gives off its own light
E F G H
It has surface temperatures as low as -200° F
E F G H
It revolves around the Sun (MAT)
E F G H

Figure 3.15. (continued on next page)

4) On Earth, day follows night and night follows day because the

A. Sun moves in an orbit around the Earth.
B. pull of the Earth's orbit affects the Sun.
C. Earth rotates on its axis.
D. tides cause the Sun to rise and set. (STG)

5) You are going to design a house so you can live on the moon. What do you need to include?

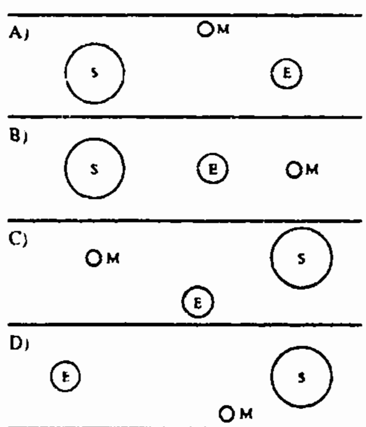
A. Food, water, books, music
B. Air, temperature control, water, food
C. Temperature control, food, air, space for company to sleep
D. Air, water, light, communication system (JEF)

Figure 3.15. Examples of Earth Science Items at the Using Level

Extending Earth Science

The items assessing the "extending" of earth science topics are displayed in Figure 3.16. In the first example the student must analyze the relationship between the Earth, Sun, and Moon, together with the fact that the Sun is the only source of direct light, to determine the positions of these bodies during an eclipse. For the second example the student must understand in what position the sun is to "see" a full moon nearly overhead and determine the time of day by the relative position of the Earth and the Sun.

1) Since eclipses take place only when the Earth, Sun, and Moon are on a straight line in space, which picture shows these bodies during an eclipse?



(STG)

2) At what time of the day would you expect to see a full moon nearly overhead?
A. Sunrise
B. Noon
C. Sunset
D. Midnight (STG)

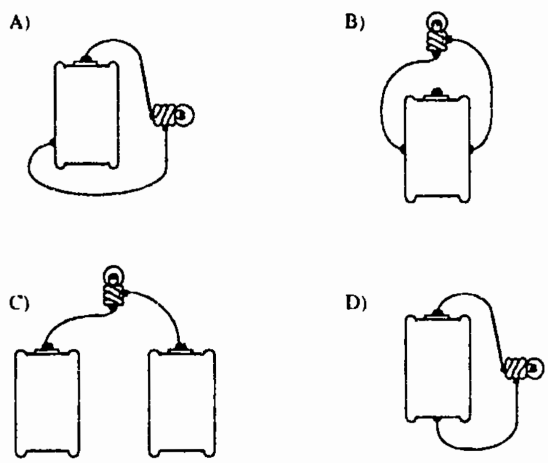
Figure 3.16 Examples of Earth Science Items at the Extending Level.

Physical Science Items

Knowing Physical Science

A number of items assessing the "knowing" of physical science topics are presented in Figures 3.17 and 3.18. The first item measures knowledge of the correct position of batteries and bulbs to make the bulb light. The second and third items test the knowledge of the properties of electrical conductivity and solubility in water. Then in items 4 and 5 of this set (Figure 3.18), knowledge of gravity and electrostatics and their effects are assessed. The following items on "water" and on "sound waves" can be grouped at the "knowing" level by the brevity of the questions and the one-word answers. In the item on "chemical change," students are told the key characteristic of chemical change and then must recall in which of the examples that kind of change occurs.

1) Which of these circuits will light the bulb?



(JEF)

2) Which list contains ONLY things that will conduct electricity?

A. salt water, steel nail, nichrome wire, copper wire
B. sandpaper, wood, water, aluminum
C. hair, rubber, steel nail, salt water
D. nichrome wire, yardstick, desk top, Coke can (JEF)

3) Which of the following does <u>not</u> dissolve in water?

A. sand
B. salt
C. soap
D. sugar
E. air (ISS)

Figure 3.17. Examples of Physical Science Items at the Knowing Level.

4) A boy threw his rubber ball into the air. Why did it come back to the ground?
   A. The air pushed it back.
   B. The Earth is a large magnet.
   C. The air is very light.
   D. The Earth pulled it back.
   E. Rubber always bounces back. (ISS)

5) What is the reason that a comb that has been rubbed can pick up piece of paper?
   A. The comb has become charged with electricity.
   B. The comb is magnetic.
   C. The comb is beginning to wear out.
   D. The comb is colder than the paper.
   E. The comb is longer than the paper. (ISS)

6) The water we drink is
   A. a solid
   B. a liquid
   C. a gas
   D. air (STE)

7) Sound waves are produced by
   A. light.
   B. heat.
   C. gravity.
   D. vibrations. (MAP)

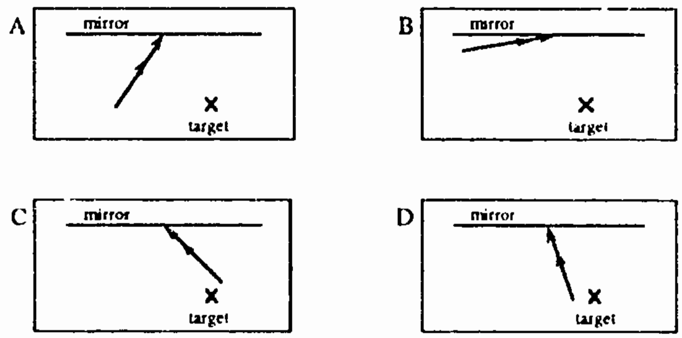
8) In a chemical change, the material you start with may be changed into something very different. In which of the following does a chemical change take place?
   A. Breaking a glass
   B. Cutting wood
   C. Burning alcohol
   D. Spilling milk (STG)

Figure 3.18. Examples of Physical Science Items at the Knowing Level.

Using Physical Science

In assessing the using of physical science topics, the first item in Figure 3.19 illustrates an application of the principle that light reflects off smooth surfaces. Students must "apply" the principle of equal angle of incidence and reflection but in a very general, non-quantitative manner. In the second item, involving the absorption of energy as a result of a change of state, the student must compare the relative "energy level" of the two states of matter in each choice. In only one choice is the second state at a higher level (e.g., gas state higher than the liquid state) and therefore a result of energy absorption by that system. The problem of the candle in different-sized containers (item 3) requires the student to compare the burning time in each of the three situations and determine the order in which each candle goes out. The relative rate of evaporation from various containers is assessed by the following item (see Figure 3.20) which requires the student to determine from which glass water disappears the quickest as a function of different amounts of surface area exposed to the air. Finally the last item, item 5, assesses the understanding of the characteristics of a series circuit: if any lights are lit, they all are lit. Only one of the four circuits has that simple characteristic of the series circuit; a circle.

1) In which case is the beam of light most likely to hit the target?



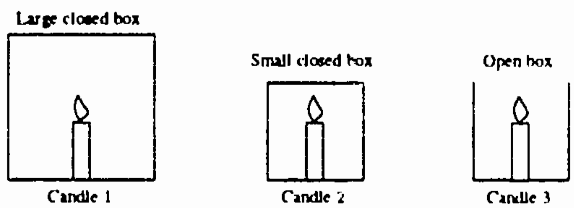
(JEF)

2) Suppose that each of the following changes of state occurs. Which of them absorbs energy?

A. Making ice from water
B. Collecting water from steam
C. Boiling water
D. Freezing fruit juice

(STG)

3) Three candles, which are exactly the same, are placed in different boxes as shown in the diagram. Each candle is lit at the same time.



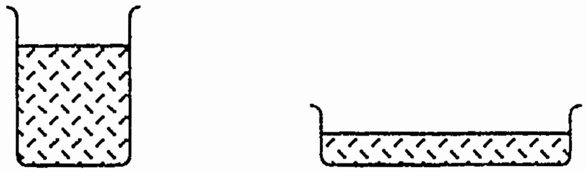
In which order are the candle flames most likely to go out?

A. 1, 2, 3
B. 2, 1, 3
C. 2, 3, 1
D. 1, 3, 2
E. 3, 2, 1

(ISS)

Figure 3.19. Examples of Physical Science Items at the Using Level

4)

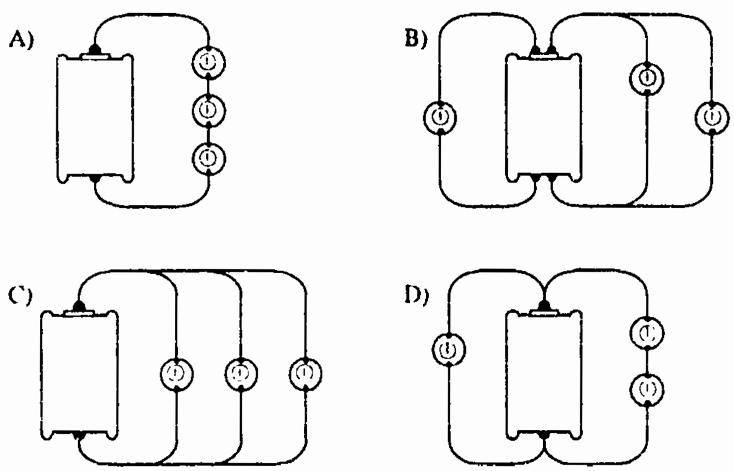


Container 1 Container 2

An equal amount of water is poured into two containers. The containers are then placed next to each other in the Sun. Later in the day there is no water in Container 2 but there is still some water in Container 1. Why did the water evaporate faster from Container 2?

A. The Sun must have been hotter in Containe
B. There was more water in Container 2.
C. There was less water in Container 2.
D. More of the water in Container 2 is closer to the air.
E. The air pressure on the water in Container 2 was greater. (ISS)

5) In which circuit could you unscrew one bulb and make all the others go out?



(JEF)

Figure 3.20. Examples of Physical Science Items at the Using Level.

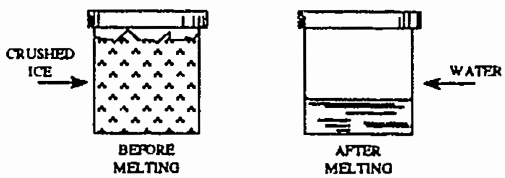
Extending Physical Science

Figures 3.21, 3.22, and 3.23 contain items illustrating the assessment of "extending" in the area of physical science topics. The "melting ice" example (item 1) presents the classic question of the conservation of mass. Students must understand that even though the ice has totally melted, all of the mass is still there in the water. This important science principle, conservation of mass, is also assessed in the second item using the weight of lumps of clay. Students are expected to know that the mass of a sample of clay is not changed when formed into smaller bits, and they must then apply this knowledge in the example. Item

3 presents the phenomenon of an expanding balloon and requires an explanation that involves the change or expansion of the air in the bottle as it is heated. This expansion then causes the balloon to get bigger. Item 4 refers to the ring of water which forms on a table from a glass of ice water and asks about the effects of condensation. It requires an analysis of the environment in which moisture condenses on cooler objects.

In the following example, item 5, the classic "see-saw" problem is illustrated. The student must analyze the mass/weight of the two people and determine the appropriate placement to "balance" the see-saw. The "ball and ramp" problem in item 6 has students compare the effects of the four different inclines on the resulting ball speed (and momentum) and therefore on the movement of the box.

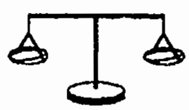
1) The can below was filled with crushed ice, sealed, and weighed. The ice was melted by slowly warming the can and its contents. No water vapor escaped, and no air entered the can.

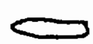


The can was then weighed again. Which one of the following results would you expect to find?

A. The weight was the same.
B. The weight was more.
C. The weight was less.
D. The weight was much less.
E. It would depend on how slowly the can was heated. (ISS)

2) Joan placed two equal lumps of clay on a balance scale as shown below.



Then she shaped one piece of clay like this: 

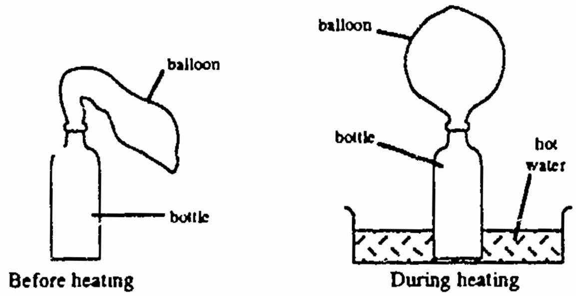
She broke the other piece into four pieces like this: 

Which picture best shows what the balance scale looked like when Joan put the clay back on the balance scale?

(JEF)

Figure 3.21. Examples of Physical Science Items at the Extending Level.

3) A balloon is placed over the mouth of a bottle. The bottle is then heated by placing it in a dish of hot water. After a short time the balloon gets bigger.

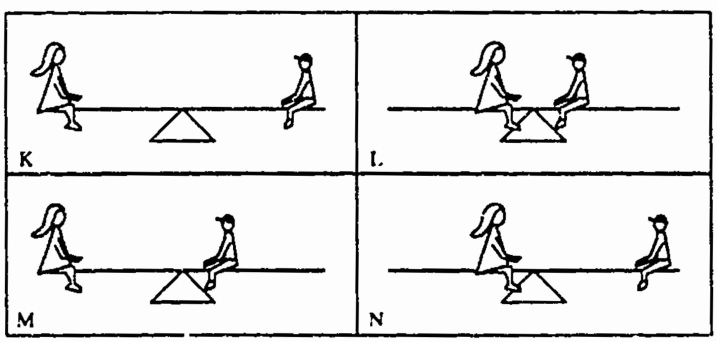


Why does the balloon get bigger?
A. The air in the bottle expands.
B. The air in the bottle contracts.
C. The air pressure inside the bottle decreases.
D. The air pressure outside the bottle decreases.
E. The glass bottle expands. (ISS)

4) A glass of ice water left on a table often leaves a ring because
A. The weight of the glass makes it sink into the wood.
B. Very small cracks in the glass allow water to leak out.
C. The ice melts and water spills over the edge of the glass.
D. Water vapor condenses on the cool outside surface of the glass. (STG)

5) A girl wants to seesaw with her little brother.

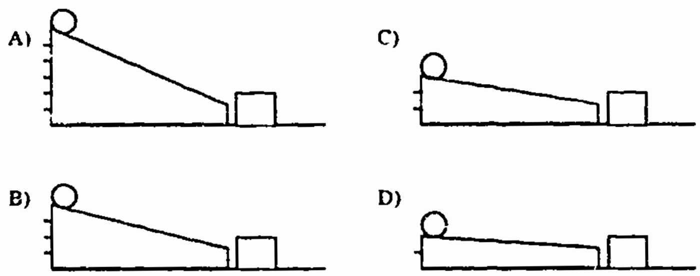
Which picture shows the best way for the girl, who weighed 50 kg (kilograms), to balance her brother, who weighed 25 kg?



A. picture K
B. picture L
C. picture M
D. picture N
E. none of these (ISS)

Figure 3.22. Examples of Physical Science Items at the Extending Level.

6) A ball is rolled down a ramp so that it bumps into a wooden block and moves it. In which case will the block move the farthest when the balls hits it?



(STG)

Figure 3.23. Examples of Physical Science Items at the Extending Level.

Life Science Items

Knowing Life Science

Figures 3.24 and 3.25 illustrate the assessment of life science topics at the knowing level. An important part of the life sciences is that set of concepts relating to the functioning of the human body, such as the lungs (see item 1). A critical skill for all sciences is that of classification. The illustrated item on "groups of living things" (item 2) is relatively difficult because all the distractor objects move and change, two of the characteristics of living things. The distractors for the item on "groups of living things" could be designed for different levels of difficulty by using other distractors, such as house, rock, books, etc. The "egg-laying animal" question (item 3) is relatively easy, except for the reversal of thinking from the presence of the word "not." Item 4 concerns a basic fact of biology: the source of a baby chick's food before it hatches.

The information about the stages of plant and animal development provides an important foundation for understanding many aspects of life science and life itself. Items assessing knowledge of these stages often involve the matching of young and adult stages of specific plants and animals, as in item 5 of Figure 3.24.

The last set of questions (items 6, 7, and 8) in Figure 3.25 tests the recognition of exemplars of plants, animals, and living things. This set could be improved by providing some negative examples.

Using Life Science

Figures 3.26 and 3.27 present items assessing the student's ability in "using life science." The first item requires the student to interpret both the tabular

Examples of Life Science Items at the Knowing Level

1) Where in the human body does oxygen move from the air to the blood?
   A. Skin
   B. Lungs
   C. Nose
   D. Kidneys (STG)

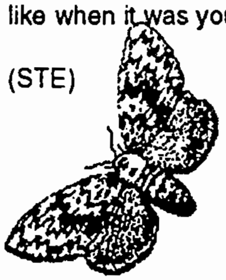
2) Which one of the following groups refers only to living things?
   A. clouds, fire, rivers
   B. fire, rivers, trees
   C. river, birds, trees
   D. birds, trees, worms
   E. trees, worms, clouds (ISS)

3) Which one of these animals does <u>not</u> lay eggs?
   A. chickens
   B. dogs
   C. frogs
   D. turtles
   E. ducks (ISS)

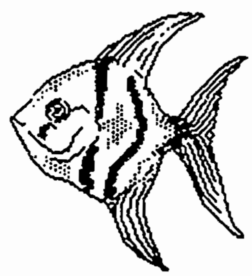
4) A baby chick grows inside an egg for 21 days before it hatches. Where does the baby chick get its food before it hatches?
   A. It is fed by the mother hen.
   B. It doesn't need any food.
   C. It makes its own food. (ISS)
   D. It uses food stored in the egg.
   E. It eats the egg shell.

5) What did this animal look like when it was young?

   (STE)



A)



B)

C)

D)

Figure 3.24. Examples of Life Science Items at the Knowing Level

6) The following questions are about the word "animal."
Is a person an animal?
a. Yes
b. No

Is a whale an animal?
a. Yes
b. No

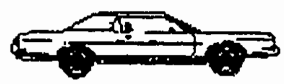
Is a spider an animal?
a. Yes
b. No

(O&F)

7) The following questions are about the word "living."
Is a fire living?
a. Yes
b. No

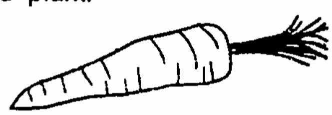
Is a person living?
a. Yes
b. No

Is a moving car living?
a. Yes
b. No



(O&F)

8) The following questions are about the word "plant."
Is a carrot a plant?
a. Yes
b. No



Is a tree a plant?
a. Yes
b. No

(O&F)

Figure 3.25. Examples of Life Science Items at the Knowing Level.

information (at the top) and the picture underneath. The student must correctly determine that ant lions prefer bright, hot, dry environments from the "chart" with the "X's." Then the student must choose the spot in the picture where the ant lions would likely be located.

Item 2 gives the pulse rates for three individuals under three different conditions. The students had to compare the range of pulse rates in these three conditions to choose which set was obtained during sleep. In item 3 students must predict what changes they would expect in pulse and breathing rates after running a race.

In biology there is a major relationship between the structure and the function of an organism's various parts. The "bird beak" example of item 4 is one way of testing that concept.

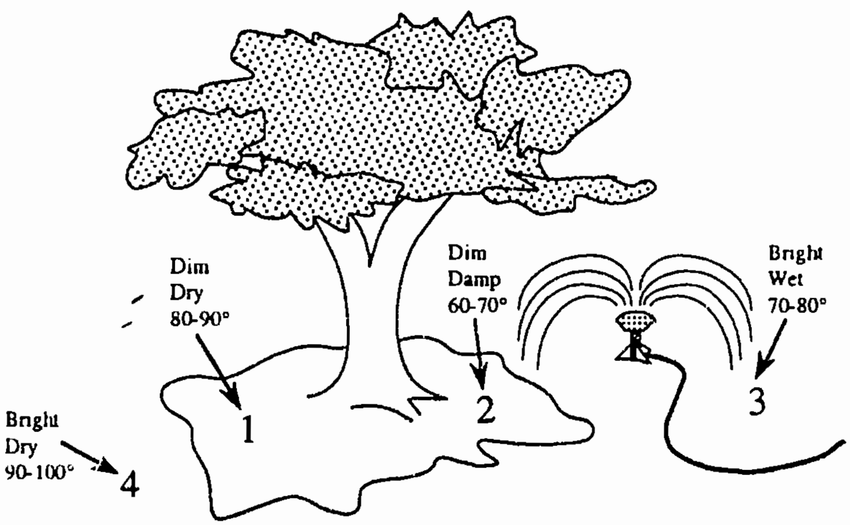
1) This chart shows the results of a test to find out the conditions that ant lions like best. The X's mark the spots where the ant lions were found.

LIGHT	dark	dim	bright
			x x x x x x

TEMPERATURE	50°	60°	70°	80°	90°	100°
					x x	x x x x

MOISTURE	dry	damp	wet
	x x x x x x		

Now look at this picture. Use the information above to find out where you would expect to find ant lions in the picture.



A. position 1
B. position 2
C. position 3
D. position 4

(JEF)

Figure 3.26. Examples of Life Science Items at the Using Level.

2) These boxes contain pulse rate data for 3 students. The data were obtained under different conditions.

Set 1		Set 2		Set 3	
Jim	60	Jim	104	Jim	68
Jane	51	Jane	93	Jane	83
John	75	John	83	John	76

Which set of data was obtained when the students were sleeping?

A. Set 1

B. Set 2

C. Set 3 (JEF)

3) Immediately before and after a 50-meter race, your pulse and breathing rates are taken. What changes would expect to find?

A. no change in pulse but a decrease in breathing rate

B. an increase in pulse but no change in breathing

C. an increase in pulse and breathing rates

D. a decrease in pulse and breathing rates

E. no change in either (ISS)

4) A boy sitting under a tree watched a bird getting insects from between the cracks of the bark. Which drawing shows the kind of beak this bird had?

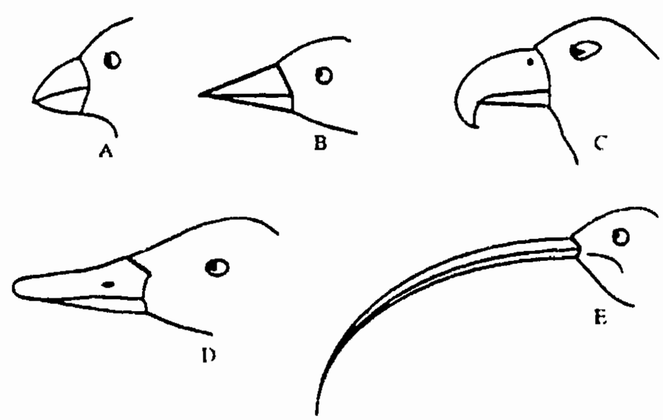


Figure 3.27. Examples of Life Science Items at the Using Level.

Extending Life Science

The first item in Figure 3.28 contains an important principle of life science: the predator-prey relationship. The "skull" item (item 1) has the student consider the characteristics (long, pointed teeth) of animals that prey on other animals for food.

The second item requires students to form a prediction, given some information about an ecological community. This item involves a forest community, and the student is asked to predict what would happen if all the plants died. The distractors relate to the components of a community, the producers, and the consumers.

The next item in Figure 3.29 (item 3) involves a sketch of several natural cycles (water, CO_2, O_2, organic matter). The student must analyze the interaction among these cycles to determine which would continue if all living things disappear.

1) A girl found the skull of an animal. She did not know what the animal was but she was sure that it preyed on other animals for its food.

   Which clue led to this conclusion?

   A. The eye sockets faced sideways.

   B. The skull was much longer than it was wide.

   C. There was a projecting ridge along the top of the skull.

   D. Four of the teeth were long and pointed.

   E. The jaws could move sideways as well as up and down.

   (ISS)

2) What would eventually happen to a forest community if all of its plants died?

   A. The consumers would have to find other producers to feed on.

   B. There would be no change in the community.

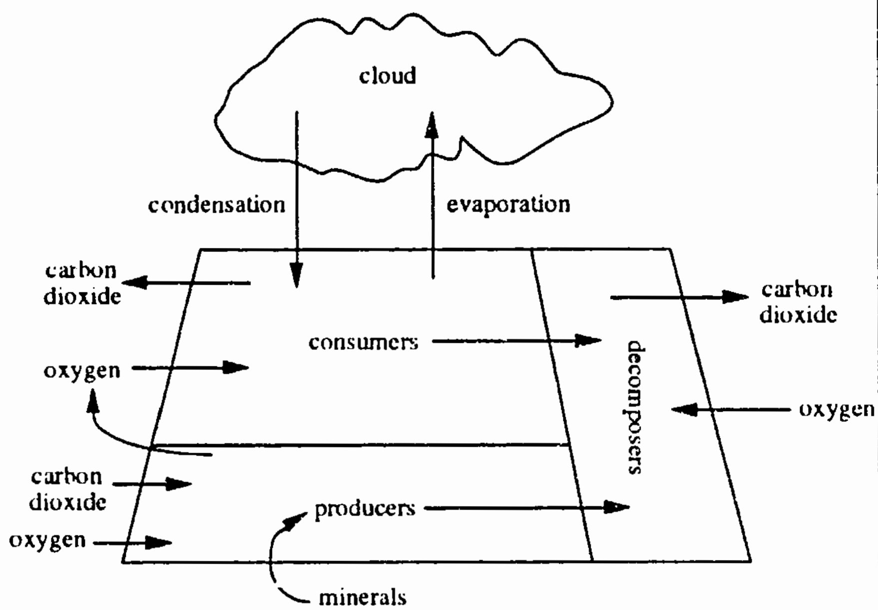
   C. The entire community would eventually die.

   D. The producers would have to find other consumers to feed on.

   (JEF)

Figure 3.28 Examples of Life Science Items at the Extending Level.

3)



A. Which cycle or cycles would continue even if all living organisms disappeared.

____

____

B. Which cycle or cycles would <u>not</u> continue if all living organisms disappeared?

____

____

(SCI)

Figure 3 29 Examples of Life Science Items at the Extending Level

CHAPTER 4

ASSESSING PROBLEM-SOLVING

Problem solving, along with the "creative/rational thinking skills" of the general curriculum, is increasingly being recognized as one of the most important aspects of science education in the elementary school. It is considered an important educational outcome, as well as an instructional method which facilitates learning of both concepts and skills. Defining or describing problem-solving becomes difficult because of such ambiguities. The term is also applied to diverse aspects of science education, such as:

- a series of steps, which may (or may not) include general and/or specific operational components

or

- mental skills involved, which may (or may not) be considered personal to the investigator and/or specific problem/situation.

However, in all contexts, problem-solving in science addresses the planning, collecting, and analyzing of data for the purpose of discovering and explaining patterns involving natural objects and phenomena. So, in the most simplistic sense, problem-solving is the application of appropriate concepts and processes to solve a science-based problem. As there is also an attitudinal component involved with problem-solving in science it can be considered to occupy the intersection of three sets (NYSED, 1985).

Models for Problem-Solving in Science

There have been a number of models suggested to describe problem-solving in the context of science. Some have been designed for instructional purposes and others for assessment. Some have been used for quite some time, while others are recent arrivals. While varying somewhat, they share certain commonalities:

- Children interact with the actual phenomena and objects (practical work).

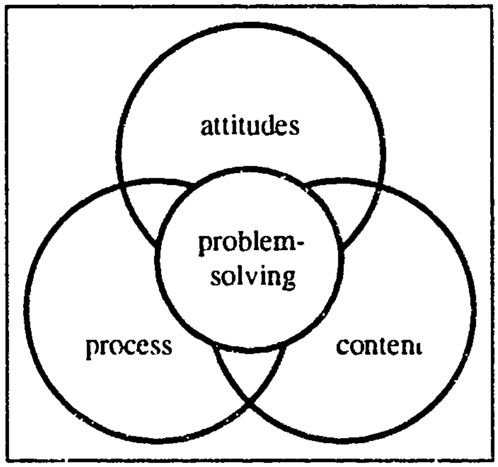


Figure 4.1. The Relationship of Problem-Solving to Attitudes, Process, and Content

- The use of higher cognitive skills is required.
- Outcomes involve knowledge (the "what" of science), as well as the skills

  (the "how" of science).
- They consist of a sequence of actions which, although appearing linear, may at times cycle back or begin at various steps.

A modified version of one of the best known, and most widely used (Lunetta & Tamir, 1979), models is illustrated in Figure 4.2 It consists of four general parts:

- Planning: the strategies for solving the problem or finding an answer.
- Performing: the operationalizing of the strategy; obtaining and analyzing the data in a "fair" manner
- Interpreting: analyzing and interpreting the data gathered; generalizing from the data
- Applying: applying the generalizations and skills to the original situation/ problem and to a similar situation or problem

Planning/design:
Question/Problem to be studied
Formulates a hypothesis
Designs a "fair test"
Performing:
Keeps test "fair"
Observes carefully
Assembles correct equipment
Measures accurately
Records observations and measurements
Performs calculations
Analyzing and Interpreting:
Transforms data into tables and graphs
Determines relationships
Formulates appropriate conclusions
Proposes generalizations or models
Explains findings
Interprets data from tables, graphs, and diagrams
Applying:
Predicts or formulates hypothesis based on the results of this investigation
Applies skills to new problem or variable

Figure 4.2. A Model of Problem-Solving

The kind of practical work that children would be involved in, in this context, is the core of the total learning process by which a solution to a problem is sought. Such an activity creates new knowledge and skill through the use of previously acquired knowledge and skills. It is more than "knowing." It requires "using" and "extending" knowledge (as discussed in the preceding chapter), as well as processes and manipulative skills. The "knowing" of various skills and concepts may be helpful but not sufficient to guarantee that our students will become proficient problem solvers (use and extend). A similar example of this type of model is given in Figure 4.3. It is taken from the elementary science syllabus for New York State (NYSED, 1985).

In the New York State guide, the program developers are reminded that before students can solve a problem, it must be recognized that there is a problem to solve. It is recommended that to help a child recognize a problem, teachers should use experiences to make them aware of discrepancies. The discrepancies arising from these experiences should lead to the raising of testable questions, and these questions will help the student to define the problem.

Questions are usually either of two types: "I wonder whether..." or "I wonder what would happen if..." The first type is concerned with a search for patterns but

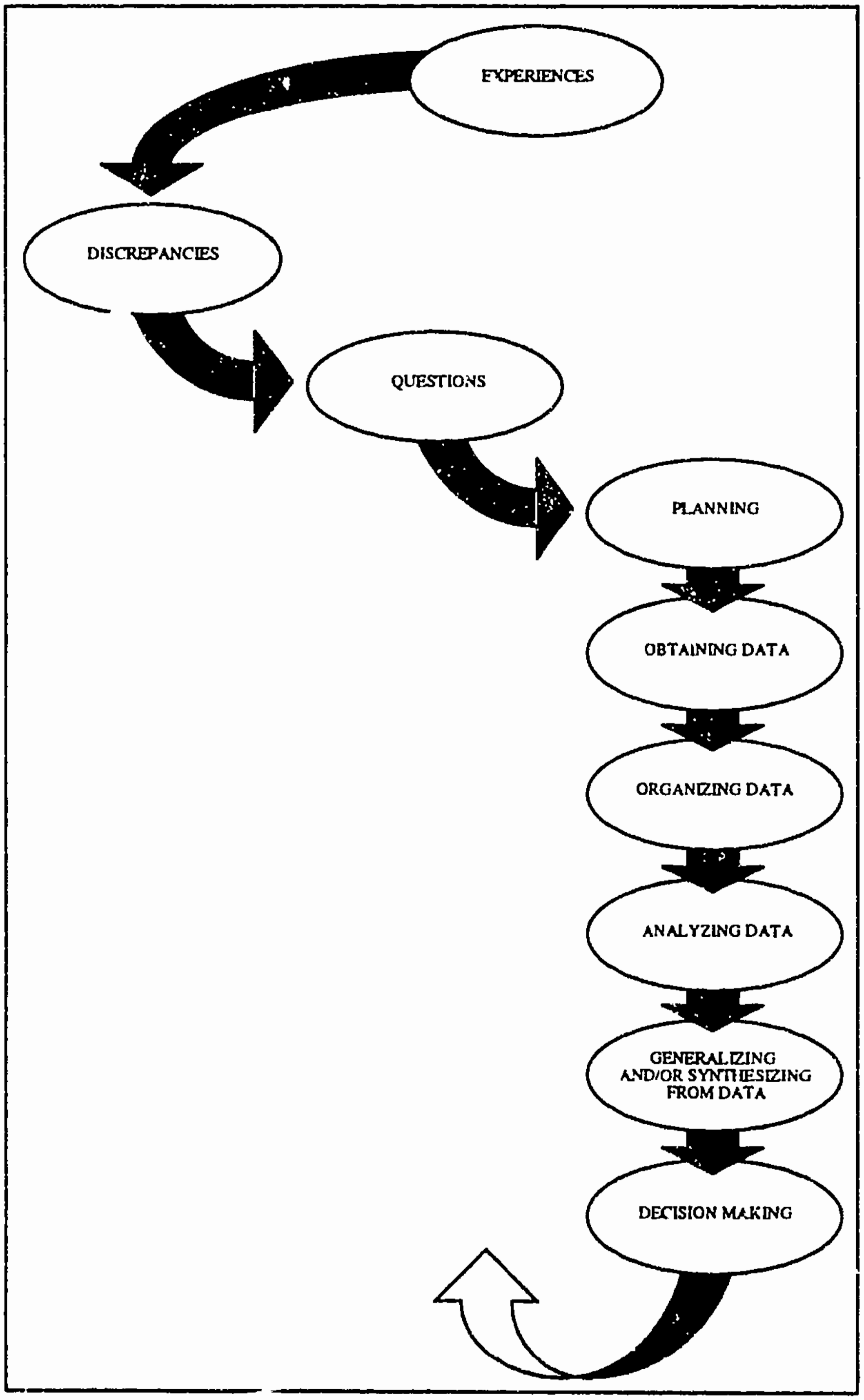


Figure 4.3 A Model of Problem Defining

often does *not* lead to an investigation in which variables are manipulated in an experimental situation. Most descriptive studies in the biological sciences fit in this category. For example, "What proportion of the population is left-handed?" or "I wonder whether pizza is still the favorite food of first graders?" The second question type, "I wonder what would happen if...?" leads to an investigation in which one variable may be manipulated in order to find an answer or solution. For example, "What effect does lowering the number of birds in an area have on the ecological balance?" or "I wonder what would happen if we increased the slope of the ramp down which we let the car roll?" Both types of questions allow for the gathering of data, whether to determine how widespread a pattern might be or to identify the effect caused by changing a variable.

Planning and Designing Investigations

While both types of questions allow for investigation, the steps in planning and performing the procedure will vary somewhat because of the inherent differences in the types of questions. The first illustration (Harlen, 1985) concerns the effect caused by changing the strength of dye used to color cloth (see Figure 4.4). It starts with a "I wonder what would happen if..." question.

Problem	How do we get the best results when dyeing?
Investigable question	What happens to the color if we change the strength of the dye?
What should be changed (the independent variable)	The amount of dye dissolved. in the investigation?
What should be kept the same?	The amount of water, the temperature, the time of soaking, the type of material, and any others which might be thought likely to make a difference.
What kind of effect should be observed? (the dependent variable)	The color.
How will the results be used to answer the question?	If there is a difference it will be possible to say what change resulted in a deeper or paler color. If not, the answer will be that changing the strength made no difference in this investigation.

Figure 4.4. Questions for Planning an Investigation

In the preceding discussion about types of questions it was stated that some types of questions *often* do not lead to investigations in which variables are controlled. The word "often" appeared because there are times when certain variables are controlled for these questions and do not result in just qualitative descriptions. A "test" is possible to determine a pattern. Figure 4.5 illustrates this type of situation (Harlen, 1985). "I wonder whether..." the hardness of wood varies in a tree?

Problem	Is the wood the same at different points up the trunk?
Investigable question	Is the wood at the bottom harder than the wood at the top?
What to look at? (dependent variable)	The hardness of the wood.
What must be different between the things looked at? (independent variable)	The part of the trunk from which blocks of wood are taken.
What must be the same? (variables to be controlled)	The direction of grain in the blocks, the position of the block in relation to the bark and the heart of the wood, etc.
How will the results be used to answer the question?	The results of the same tests on both blocks must be compared to see if the block from the bottom was harder than the block from the top.

Figure 4.5. Questions for Planning an Investigation.

While independent, dependent, and controlled variables are included in both illustrations, the types of questions in the steps differ.[1] For example, the question, "What should be changed in the investigation?" becomes the question, "What must be different between the things looked at?" in the second illustration.

"Doing" the Investigation

The actual "doing" of the investigation involves a variety of interacting manipulative and mental skills. It is a complex operation not divorced from content and, although separable for analysis and discussion, not separate in practice. Like the carpenter who knows how to use tools, can plan a project, and can judge the quality of another's work, he is not considered "expert" until he expresses his skill in the actual production of a piece of furniture. The familiar saying, "The whole is more than the sum of its parts," certainly applies to problem-solving in the context of doing an investigation in science.

During the performance of an investigation many of the "how" skills are used: how to make both qualitative and quantitative observations; how to manipulate apparatus; and how to record, graph, describe, and make numerical calculations. Judgments and decisions are made about what tool to use, which observations are relevant to the investigation, what record organization will be appropriate to make the pattern clear, and so on. The actual performance of an investigation designed by the student requires more than the ability to follow directions on a lab sheet and to come to a predetermined answer. Because an experiment does not always proceed according to our preconceived plans, the unanticipated has to be managed during the actual performance of the investigation. These adjustments, or "mid-course corrections," require judgments and decisions to be made in order to work toward formulating an appropriate conclusion or a solution to a problem. For example, when it is found that the plants in the light dry up more than the plants in the dark, a change from giving both the same amount of water has to be made in order to keep the test "fair." Otherwise, invalid conclusions can be drawn about the effect of light on plants. The performance of an investigation, although executed through the manipulation of materials and apparatus, is not just a manipulative experience composed of a sequence of steps that are invariable.

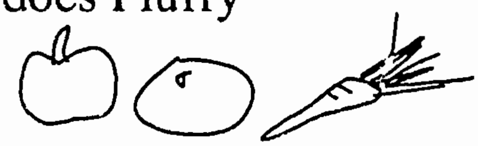
Figure 4.6 is a record of what a first grade class did in order to determine which of three foods their guinea pig preferred (Griffith-Miles, 1989). The teacher led a class discussion for the purpose of planning a test that was "fair," but changes still had to be made as they proceeded. For example, a first attempt failed because the guinea pig was not hungry enough to be interested in the food that was offered; so they decided to remove the food for a few hours before trying again. A flat bowl was used when they realized that the animal did not have a real choice of food if many of the pieces were buried under other pieces. They gave up the idea of using lettuce when they could not decide how to measure it. Decisions had to be made about what kind of record to keep, what kind of observation to make, and for how long. Many skills were used in designing the investigation, implementing the plan, and reaching conclusions based on the data

Analysis and Interpretation

The analysis and interpretation of the data gathered during the performance of the investigation are considered the most cognitive aspects of investigating. It is concerned with what is done with the data in order to uncover relationships and patterns pertinent to the original problem or question. While the skillful performance of an investigation is vital for scientific investigation, it is of little use without the perceptive analysis and interpretation of the data. For example, Tycho Brahe was a meticulous observer of the motion of the heavenly bodies and their motion in the universe. However, it took the ability of his student, Johannes Kepler, to analyze and interpret the data so that the orbital motion of the planets around the sun could be theorized. Some of the other less than elegant

An experience chart made by a primary class to describe their fair test. Fluffy, the guinea pig, was given a choice of foods. The data were collected and recorded on graphs. Their summary and conclusions were written individually. (Fluffy "liked" apples best.)

Which food does Fluffy like best?



1 To make it a fair test, we will cut the apples, carrots, and oranges into three equal parts.
2 The food will be the same length.
3 We will take three of each food.
4 We will take the guinea pig food out of the tank.
5 We will put the fruits and vegetables in a clean bowl.
6 We decided to stop the experiment when she eats all of one food.

This is what Fluffy did.

1 First she took an orange. She nibbled the orange.
2 Next she sampled the apple. She ate one of the apples.
3 Last she nibbled on a carrot.
4 We observed her for 30 minutes.

Figure 4.6. Example of Student-Recorded Data

interpretations of planetary motion now seem such unlikely candidates for a theory that it is hard to understand how they could once have been believed. (But that is because we have the benefit of hindsight.) The conclusions and generalizations that we seek grow out of the careful analysis and interpretation of our data.

Children can also come to faulty conclusions because of poor analysis and interpretation of data. A class that had just finished an investigation of various insulating materials was guided by the teacher to question why white clothes were usually worn in summer and dark clothes in winter. The teacher wanted them to learn from this next experience (dark and light clothes) that some colors absorb heat and others reflect it. She went through all the appropriate steps in planning and designing the investigation and the performance of it. When it was time to analyze and interpret the data, the students graphed the data, determined quantitative and qualitative relationships, and discussed the accuracy of their data (whether their test was fair, etc.). However, as one might guess, the children concluded that some colors *insulated* better than others. The children had over-generalized from a limited investigation. When they tried to apply this generalization to a situation in which light was not involved, they were forced to go back and reinterpret their data.

Application

Our role as teachers is to help children learn skills, content, and attitudes that resist forgetting and that will transfer to new situations. While there may be a few things that we learn just for the sake of learning, the bulk of what is taught is generally assumed to be of some useful purpose. By applying what we have learned, we clarify and refine our ideas and skills.

It is through application that we determine whether a child can "use" the skills and knowledge that we help children to learn. The application phase of investigating involves making predictions based on the conclusions drawn from the investigation and formulating hypotheses about the same or related situations. It also includes applying the techniques used in investigation to new problems and investigations. From an investigation of the behavior of a pendulum, for example, a student will make predictions and hypotheses concerning the calibration of a grandfather clock. The student would also realize that more than one try is needed to check on the accurate timing of the clock. (Even young children, who have had experience with fair tests, can understand why it is often necessary to drop a set of balls more than once in order to conclude which one bounced the highest.) Application of both content and skills are reasonable expectations of problem-solving within the context of practical science investigations.

Student Variables Involved in the Ability to Plan and Perform Investigations

There is a big step from the general strategy or design of an investigation and the actual "doing" of it. The steps in the plan must be operationalized and made more specific. For example, the actual "how" of keeping everything the same except the one variable to be investigated has to be considered, as well as the choosing of the appropriate materials and equipment for the activity. For the younger, elementary-school-aged children this is a difficult task. Most often a child, starting with only the vaguest idea or strategy of how to begin, will proceed to investigate, making mid-course corrections as needed along the way. The difficulty occurs because children have trouble anticipating, in thought, the results of a plan of action unless the type of problem and the procedure to solve it are very familiar to them. Thus, their thought is closely tied to "doing." Experience with investigating will increase skill up to a point, but mastery awaits further natural cognitive development.

For the youngest children other factors are also involved. The original question must be kept in mind, as well as the starting conditions. The ability to deal with more than one variable and with abstractions is also influenced by development. Thus, we do "fair tests" with young children, not controlled experiments with multiple variables. Even the ability to perceive discrepancies may be limited. Young children do not seem to think there is any discrepancy in the moon "following" them while also seeming to follow another child. This should not be taken to mean that assessment, or instruction, should "wait" until the child is ready; only that we must be aware of what we are really assessing. Indeed, it is important to determine where a child is, due to both experience and maturation, in order to plan for future learning experiences.

There has been increased interest in the identification of levels of ability pertaining to the skills used for investigating (and therefore problem solving). Many of the results, including the one illustrated in Figure 4.7, show evidence of a strong Piagetian influence (Shayer & Adey, 1981). As this work is in terms of descriptors of student behaviors (what a child can "do"), it holds potential for classroom use that could be of great value to the teacher. A child's ability to control and exclude variables at the concrete level, for example, (see Functions 2.5 and 2.6 of Figure 4.7) would indicate that a "fair test" is the appropriate form for instruction, not the controlled experiment. (However, this might be introduced for those who seem almost ready to deal with it.) The concrete operational child's need to deal with concrete referents is also indicated by Functions 1.1 and 1.2 of Figure 4.7.

The value of such indicators, or benchmarks, for assessment lies in the potential for indicating what one could reasonably expect a child to learn from instruction, as well as the level to which the skill has been developed. It can also be used to infer the type of assessment situation, or test format, in which the data for assessment can best be gathered. For example, most elementary-school-aged children (concrete operational children):

- have a limited ability to plan and design investigations on paper,
- need concrete materials to tackle a problem,
- need problem situations that are familiar, and
- need problems that are not abstract or complex.

It would follow that these constraints would also apply to assessment situations.

Function	1 pre-operational	2A early concrete	2B late concrete	3A early formal	3B late formal
1.1 Interest and Investigation style	Things are believed to be exactly as they appear to immediate perception. Perception dictates decisions. Faced with a mature person's idea of evidence, will deny it, explain it anthropomorphically, or be silent. Does not perceive contradictions.	Will register what happens, but for interest to be maintained after the first obvious observations needs a serative or simple associative model[4]. Unaided investigation style does not go as far as producing concrete models.	Will include seriation[1] and classification[2] as tools of perception in finding out what happens, but needs to be provided with a concrete model by which to structure experimental results (classes must be given, and examples of the application shown). Finds interest in making and checking cause-and-effect predictions.	Finds further interest in beginning to look for *why*, and following out consequences from a formal model. Confused by the request to investigate empirical relationships without an interpretive model. Can use a formal model, but requires it to be provided. Can generate concrete models with interest. Can see the point of making hypotheses, and can plan simple controlled experiments but is likely to need help in deducing relationships from results and in organizing the information so that irrelevant variables are excluded at each step.	Finds interest in generating and checking possible 'why' explanations. Will tolerate absence of an interpretative model while investigating empirical relationships. Takes it as obvious that in a system with several variables he must 'hold all other things equal' while varying one at a time, and can plan such investigations and interpret results. Will make quantitative checks involving proportionality relationships.
1.2 Reasons for events	Interprets phenomena egocentrically, in terms of his own self.	See 1.1 and 1.3: Cause-and-effect only partly structured—'this goes with that'; so uses associative reasoning. Simple one-factor causes, such as 'force', etc.	Bipolar concepts such as 'alkali destroys acid'. Can use ordering relationships to partially quantify associative relationships: 'as this goes up, that goes down', 'if you double this, you must double that; i. e., 'the reason' involves describing the relationship or categories, not providing a formal model. Cause-and-effect structured according to general concrete stage schemas as 'adding acid makes the pH lower.'	Looks for some causative necessity behind a relation established with concrete schemas. Allows for the possibility of a cause that is not in 1:1 correspondence with observations. Can consider the possibility of multiple causes for one effect, or multiple effects of one cause. Can suspend judgment and allow results of controlled experiments to constrain choice among various cause-and-effect explanations. Can handle formal model as explanatory provided their structure is simple.	Becomes aware of multiple causes and effects, can think of reality in a multivariate way, so can make a general or abstract formulation of a relationship which covers all cases in an economical way. Can use deduction from the properties of a formal model—either from its mathematical or internal physical structure—to make explanatory predictions about reality.
1.3 Relationships	Cannot consistently arrange data in an ordered series	Can order a series, but is unlikely to see that as an obvious way of summing up observations. Nominal scale[3] relationships—'same distance - same weight' (see-saw).	Can multiple seriations, and hence, can find 1:1 correspondence between two sets of readings (e.g., weights and extensions of springs), and hence any linear relationship. Readily uses the notion of reversibility. Will use compensation argument to explain a conservation where only ONE variable is independent, e.g., of a piece of clay 'you've made it longer, but it's thinner so it's the same.'	Uses compensation relationships between full independent variables, e.g., weights and distances on both arms of a balance can be changed while preserving equilibrium; resistance is related to both area and length, in electricity. Simple functional relationships beyond linear, and thus acceleration. This is a more sophisticated version of concrete modeling, rather than formal modeling.	Can reflect upon reciprocal relationship between several variables. Thus can handle quantitative 'relations between relations' as in proportions, or semi-quantitative relationships as in chemical equilibria. This level of thinking is often needed for analyzing experimental results so as to order them for lower-level computation, e.g., weight changes in reactions involving different elements and compounds, or density calculations where density is an inferred concept (density of gases, or Archimedes' problems).

1 Seriation: putting objects or data into order according to a property such as length or mass (ordinal scales)
2 Classification: putting things in groups according to common property.
3 Nominal Scale: scale with two values only - 'Four legs good, two legs bad'.
4 Association: coincidence in time or place serving as basis for prediction.

Figure 4.7a. Examples of Problem-Solving Skills by Development Levels

Type of Problem	2A early concrete	2B late concrete	3A early formal	3B late formal
2.5 Control of Variables	Can reject a proposed experimental test where a factor whose effect is intuitively obvious is uncontrolled, at the level of 'that's not fair', but fails to separate variables and so eliminate one. 'Fairness' may also be applied in the sense of giving every factor an equal chance, e.g., 'Slower runner should be given a shorter distance.'	Will usually vary more than one factor in each experiment, and often varies other factors to test the effect of a given one.	Sees the need to vary one factor at a time and can suggest experimental tests to control for factors explicitly named. May fail to control factors that are not perceptually obvious. Fails to develop a strategy based on a feeling of the system as a whole. May not see the point of having an experiment without a factor present to see if it is a variable.	Sets up suitable experiments to economically control factors and eliminate ones that are not effective, and can apply 'all other things equal' strategy to multivariate problems. More sophisticated biological experiments possible including interaction effects. Appreciates the impossibility of controlling natural variation, and so the need for proper sampling.
2.6 Exclusion of Variables	In analysis of a multivariate problem (e.g., pendulum, flexibility of rods of different materials, shape, etc.) has no strategy for excluding interfering variables. May attempt to order the effects of factors and may arrive at the direction of the effects if they are intuitively obvious, e.g., 'longer rods bend more.	Will order the effects of a given factor, but fails to exclude the interference of the other factors because he is trying to impose bivariate thinking. Thus often arrives at correct effect of a factor by invalid arguments. Unlikely to arrive at correct effect where it is contrary to intuition or where the factor makes no difference.	Will correctly arrive at the effect of a factor from experiments in which he/she has controlled for the other factors, but will often fail to exclude the effect of other factors when asked to select, from a group of experiments, those required to show the effect of each factor. Thus when two factors have been changed, and an effect is noticed, is likely to attribute change to the combination of both.	Because of an implicit knowledge of the different effects which may be caused by the combinations of the variables that are possible will select economically from a variety of experiments those required to show the effect or non-effect of each in turn.
2.7 Probabilistic Thinking	No notions of probability.	Given 3 red objects and 3 yellow objects mixed up in a bag, realizes that there is a 50/50 chance of drawing a red one.	Given other ratios of objects will count the numbers of the given type (n) and the number of all objects (N) and express the chance of selection as a fraction n/N.	
2.8 Correlational Reasoning		No systemic method of estimating the strength of a relationship except to look to see if the confirming cases are bigger in number than all the rest.	Begins to look at the ratio of confirming to disconfirming cases, but tends to look only at the probability of two of the four cases, e.g., for blue or brown eyes and light or dark hair will compare the ratio of those with blue eyes and blond hair to those with blue and dark.	Realizes that the opposite pairs are as important as each other. Thus takes the brown hair/dark eyes set together with the blue eyes/fair hair set, and compares it with the sum of the two disconfirming cases (brown/fair and blue/dark).
2.9 Measurement Skills	Makes measurements by comparing beginning an ending of object/journey with rule in simple whole numbers	Bar diagrams, histograms, idea of mean as the center of a histogram, and variation as its breadth. Graphical relationships of first order equations. Interpretation of graphs where there is a 1:1 correspondence with the object measured, e.g., height/time relationship for the growth of a plant	Interpretation of higher order graphical relations, and use of problem-solving algorithms, e.g., $P_1V_1=P_2V_2$ for gas pressure calculations. Can make interpretations which involve relations *between* variables in a graph, e.g., in a distance/time graph will see that a horizontal section means 'standing still' and that a vertical section is impossible	Interpretation of higher order graphical relations in terms of *rates* (instantaneous slopes) and reciprocal relationships; conceptualization of relationships between variables, e.g., in $V=lbh$, if l rises (V constant), b and/or h must drop proportionally

Figure 4.7b. Examples of Problem-Solving Skills by Development Levels (Continued)

Benchmarks Along the Way of Progress

It is helpful to have benchmarks along the way to mastery of the various objectives of science. They give us an indication of where we are and the direction in which we are going. By looking back we can see what has already been accomplished. It allows us to assess where people are along the journey and help with the next steps in the process. There have been some attempts to establish steps along the way. Some were illustrated in Chapter 2 on process skills. The same sources are used for an illustration of some of the skills involved in problem solving and investigating in this chapter.

The first illustration, shown in Figure 4.8, comes from a *State Guide to Science Curriculum Development* (Wisconsin Dept of Public Instruction, 1970). It spells out a series of behaviors from "a" through "j." There are more items included in this scheme than the next scheme that will be illustrated. The behaviors span the interval of concrete operations, with the last few being more appropriate for older students who have had considerable experience and are transitional or capable of formal operational thought.

Process Sequence

a. Manipulating apparatus to make pertinent observations

b. Identifying observations which are relevant to an experiment

c. Distinguishing useful from extraneous data

d. Describing the problems involved in making desired observations

e. Identifying the relevant variables in an experimental situation

f. Maintaining an accurate record of experimental procedures and results

g. Controlling those variables not a part of the hypothesis being tested

h. Identifying sources of experimental error

i. Describing the limitations of experimental apparatus

j. Describing the limitations of the experimental design

Figure 4.8 Process Skills for Experimenting

The next two illustrations are checklists intended to be used during normal activity in the classroom (Harlen, Black & Johnson, 1981). The first set, shown in Figure 4.9, was taken from a group designed to be used for younger children. The second set, shown Figure 4.10, is intended for use with older elementary-school-aged children. Some overlap occurs between the two sets. They were designed to reflect the interaction of natural maturity and learning. For example, in the middle box for "Problem Solving," the skill in determining which way of tackling a problem is relevant, is a function of both experience and maturation. The problem of determining relevance can also be inferred from the middle box listed under "Applying Learning." The second and fourth boxes are furnished to indicate transitional levels between the three levels with detailed descriptions. They are intended for assessment through teacher observation of students' behavior as they are involved in practical activities.

Levels of Mastery for Problem Solving

Generally unable to approach a problem without help.	Tries one or more ways of tackling a problem without much forethought as to which is likely to be relevant.	Identifies the various steps which have to be taken and tries to work through them systematically.

Figure 4.9 Levels of Mastery for Problem Solving

Levels of Mastery for Applying Learning

Rarely connects previous learning with new situations in which it could be applied unless told what skill or idea is relevant.	Uses previous experience in new situations once the relationship between the new and previous situation has been pointed out.	Works out what earlier learning could be applied in a new context by using relationships between one situation and another.

Figure 4.10 Levels of Mastery for Applying Learning

Assessment of Problem-Solving in Science

Problem-solving in science usually takes place in the context of an investigation, or "practical" science. A recent review of assessment in science (Bryce & Robertson, 1985) states, however, that "the bulk of science assessment is traditionally non-practical." This condition still exists in spite of the wide-spread belief of the importance of practical science. There are potential problems if assessment of problem-solving outcomes is attempted by *only* paper and pencil methods, some of which are summarized below:

- The danger of the test becoming the curriculum.
- Evidence that seems to indicate that non-practical science does not require the same mental processes as those used for practical science.
- The nature of the concrete operational children and their strengths as well as limitations.
- The indication that problem solving is more than just the sum of the skills and knowledge involved. It is a holistic task.

A "best-guess" about student ability to solve problems would ideally be made from observations of students' actual performance in problem situations which were made over a period of time. However, a paper and pencil test can be more conserving of resources, both human and otherwise, than observational assessment methods. As so often happens in the educational enterprise, we are faced with a dilemma that requires that some decisions and judgments be made before the original task can be accomplished.

We can allow ourselves a wider range of assessment methods than just observation of student performance if we look at the task as analogous to putting a puzzle together to complete a picture. The choice of using only paper and pencil testing, however, is no choice at all if we claim to be measuring the problem-solving objectives of our programs, for assessment of only a *segment* of the objectives can not stand for a valid appraisal of the whole program. Therefore, at least part of the assessment picture must be a practical component, while other parts may take different forms. One could also argue that an assessment system of just practical assessment might overlook some important outcomes.

There are ways of combining both practical and written tasks in order to obtain a written record of children's work that allows us to determine, at a later time than when the test was given, how well the children can manage a problem-solving situation. Decisions have to be made about which segments are more appropriately tested by practical means and which lend themselves to paper and pencil assessment. For example, those listed under "Performing" in Figure 4.2 would be better assessed in a practical situation, although a record written by the student of some of his work could be used. It must be kept in mind that changing the method of assessing may change the task being required of the student. For example, asking students to construct a fair test and phrase it in their own words is different than either requiring them to follow and judge the soundness of someone else's plan or requiring them to demonstrate this skill in a practical mode.

Practical assessment of problem-solving skills can be done through either of two basic ways: by teacher observation, supplemented with questions to clarify what the student is, or was, doing, or by a "lab test" method in which students read or are told the task, perform the task(s), and then respond in written form. Teacher observation of students is better suited for individual assessment, whereas the lab test or "group practical" can be administered to a group of students who all have the same assignment (and materials) or who rotate around to stations containing different tasks

Teacher observation of an individual child allows for assessment of a wide variety of skills involved with problem-solving in practical science. For example, although the child may not be asked to record a strategy for investigating before starting the activity, the operationalization of a strategy can be observed, as well as "mid-course" corrections as he/she proceeds through the task. During the actual gathering of the data, the student will demonstrate the degree of skill in doing the test fairly, observing carefully, and measuring accurately. The student is also required to know how to use the equipment, recognize pertinent data, and use symbols appropriate to the particular investigation. During the analysis and interpretation of the data, the student acts on the data to identify relationships and come to conclusions consistent with the findings of the investigation.

Interaction with the tester is more limited in a group practical situation. The administrator of the test needs less training and the method yields a permanent record that can be marked at a later time. The tasks are usually shorter, done within a set time limit, and assess a more limited number of skills. For example, although a child might be required to conduct a fair test, only the data and/or conclusions may be recorded in the student record booklet. How the student went about the test, which an individual practical situation could assess, can only be inferred. While more information about student performance *could* be gathered, the amount of reading and writing would probably not be appropriate for elementary-school-aged children. (As the questions and answers are usually in written form, some ability in this area is required. This may serve to penalize some children.) No hints are provided by the tester to a student if he gets stuck and can not proceed.

A set of three items may comprise the test, with each item assessing somewhat different skills. Materials to be used for each task are set out at stations. While more students can be tested in a group practical per given time interval than in an individual practical, more materials are needed to accomplish this. Although items are selected that can be finished in approximately the same time, individuals will vary in the time needed for each. Therefore, while the group practical may save time, the method has disadvantages that the individual practical does not.

The following items have been chosen to illustrate some of the tasks used to assess problem-solving in science in the elementary school:

Bouncing Balls - an individual practical

Bouncing balls is considered an experience common to children, so the focus of assessment in this task is on problem solving skills, while the concepts that are involved are ignored. A tester introduces the task to the student and records what the child does. A student record sheet is provided which repeats the problem that was orally given to the child and includes a place to record data and conclusions. What is included on the sheet serves as a reminder of what is to be done but not how to do the task. Materials are provided to the child. (These could, of course, serve to limit the investigation.) This item was used to test both 11 and 13-year-olds. Eleven-year-olds were observed during performance. Thirteen-year-olds wrote their plan but did not perform the task (Harlen, Black & Johnson, 1981).

Student Page for Bouncing Balls

Bouncing Balls Pupil Number ____

Find out if the ball which bounces best on one surface also bounces the best on all the surfaces

a) Put down here any results and working as you go along:

b) Put down here what you found out:

Surface	Best bouncer
Carpet	
Rubber	
Ceiling tile	

c) Does the ball which bounces best on one surface bounce the best on all surfaces?

Figure 4.11. Student Page for Bouncing Balls.

Circuits - an individual practical

This item is similar to the preceding one in that the tester presents a question which the student attempts to answer as the tester observes and records what was done (Harlen, Black & Johnson, 1981). A student record sheet is included. Hints, in case the student has trouble proceeding, are provided as well as a question intended to elicit from the student an evaluation of his experimental procedure. For example, the child's attention can be directed to the extra wire, if he does not seem to realize it is needed. "How would you do it next time?" would be asked after the student completed the task.

A complete circuit is given for the child to inspect and to light the bulb by pushing the button switch. Then an incomplete circuit with two faulty parts is given to the child. He is asked to make the bulb light, in this second circuit, and then explain why it had not lit as originally given to him. This activity is more demanding of the child than "Bouncing Balls" and would take longer to finish.

While the student needs to use a procedure to complete this task, it is somewhat different than that in the item about bouncing balls. It would consist of the systematic elimination of a potentially relevant variable when it was found not to effect a change (i.e., the bulb would still not light). Since there are two "faulty" items in the circuit, it is a complex problem.

The content of the task is not considered an experience common to children, as was the item dealing with bouncing balls (it depends on the learning of school science). For this task the student would compare the arrangement of materials with the model provided and make a series of inferences based on the differences and similarities. Experience with these materials could give a child an advantage over a child who had not worked with them previously. It would make the task more familiar to him.

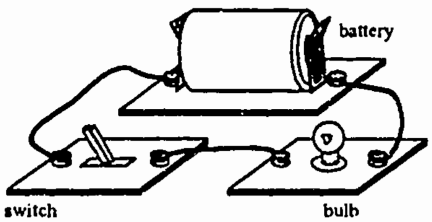
Water-Temperature Activity — an individual practical item

This is not an "experiment" in the same sense that the two preceding illustrations could be called experiments (NAEP, 1975). This is an activity in which the student follows oral directions and gives oral responses to the tester's questions throughout the duration of the task. This task was given to a selected group of children ages 9, 13, and 17. There would seem to be an aspect of this task that is dependent on the level of cognitive development of the student. This task was also used as an item on the SISS group process test, however in totally written form. There was a chance, therefore, that the child could fill in the "predicted" temperature of the mixture after having mixed the two quantities. The student is provided help in using the thermometer if he/she does not have this skill. A simple "workbook" page is given to the child to record the actual temperature, as read from the thermometer, and the predicted temperature of the mixture.

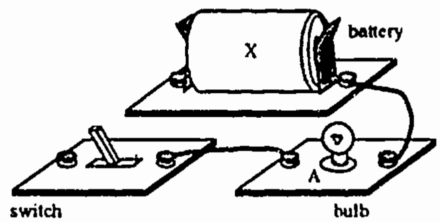
The task consists of reading the temperature of two equal amounts of water, one of which has been heated. The child is then asked to predict the temperature of the mixture of the two and is then instructed to mix the two quantities in a third cup and read the temperature of the mixture.

Example of an Individual Practical Test

Materials:
A mounted battery, bulb and switch connected in a circuit as indicated in the diagram.



A dead battery, marked X; bulb in holder, marked A; and switch connected as indicated in the diagram.



Also:

4 spare wires with connecters at both ends;

Spare battery, marked Y, in holder;

Spare bulb in holder, marked B;

Electric bell.

Student Record Page:

Circuits

Try to make the bulb light. When you have made it light, find out what was stopping it from lighting in the beginning.

a) Put down here any notes about what happens when you try different things:

b) Put down here why the bulb did not light in the beginning:

Figure 4.12. Example of an Individual Practical Test

The administrator of the test is given a set of directions telling, step by step, what to do and say to the child. He records such things as whether the child was given help reading the beginning temperatures, the actual temperatures of the water when he checked it subsequent to the student reading, and the student's prediction of the temperature of the mixture. The student is not asked to explain any discrepancies between the predicted value and the actual value that was determined by mixing the two quantities of water.

Although the student is conducting a simple test in this activity, it is at the direction of the adult involved. Therefore, only skill in predicting and use of the thermometer seems to be assessed from this task. (Results were given for "Do you know how to use a thermometer," and for "Verbally explaining the results of mixing the water.")

Administrator Instructions for Water-Temperature Activity

(Take out three styrofoam cups, two plastic measuring cups, two Centigrade thermometers, a plastic bottle containing cold water (the water in the bottle should be as cold as is readily available), a water heater, and a few paper towels. Place all of the materials in front of you.)

(Measure out 50 milliliters of the cold water into each of the two measuring cups.) I am measuring out 50 milliliters of cold water into each of these two measuring cups. You see the water is equal. (Indicate to respondent that the measuring cups contain the same amount.)

(Pour the water from the measuring cups into two of the styrofoam cups.) Now, I will pour the water into each of your cups, and heat the water in the one labeled HOT until it is quite hot. (Place the water heater into the cup labeled HOT and heat a few seconds until water is hot to the touch.)

(Place the two cups before respondent.) The two cups are for you to work with. Remember, the one labeled HOT contains a certain amount of hot water and the one labeled COLD contains the SAME AMOUNT of cold water.

(Give respondent the two thermometers and the Workbook opened to the correct page.)

A. I have also given you two thermometers. Do you know how to use a thermometer?

- ☐ YES (Go to C)
- ☐ NO (Go to B)
- ☐ I don't know. (Go to B)
- ☐ No response (Go to B)

B. (Show respondent how to use a thermometer. Include: How to hold carefully, insert in water so as not to tip the cup, wait a period of time, and read.)

C. (Put a thermometer into each cup of water and record the temperatures in Part A and Part B.)

(After respondent reads and records the temperatures of the hot and cold water, read and record your readings below. If the discrepancy between your readings and respondent's readings exceeds two degrees, help respondent to read the thermometers and have him record correct readings.)

Figure 4.13a. Administrator Instructions for Water Temperature.

Workbook Page for Water-Temperature Activity

A. Temperature of hot water is ____________° C.

B. Temperature of cold water is ____________° C.

C. I THINK temperature of the mixture will be ____________° C.

D. Temperature of the mixture is ____________° C.

Use the space below to do any work:

Figure 4.13b. Workbook Page for Water-Temperature Activity.

A Test for Electrical Conductors — a station practical item

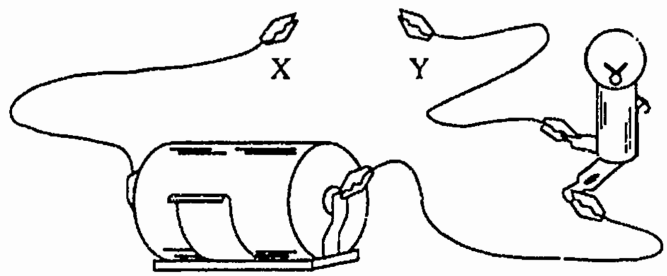
This item is from the SISS test composed of three practical tasks (SISS, 1986). The student is not required to plan a fair test or investigation, so it is restricted in the variety of skills assessed. It requires the student to read and follow directions and to construct an electrical "tester" from a picture of one. In order to complete the task, the student needs to know (or infer) what the term "conductor" means.

The student is given materials from which he puts together the tester. After trying the tester to make sure the bulb will light (it is functioning as it should), he is then instructed to determine which of the materials in a set of materials light the bulb and which do not. From this he determines which conduct electricity. Lastly, he has to give reasons in his own words why he concluded from the data which objects were conductors of electricity.

As with the other item involving electricity, the content of the task is not considered an experience common to all children. The student collects, analyzes and interprets the data, and comes to conclusions based on the data that were gathered. However, experience with this activity, such as school instruction, would give the student an advantage over those who did not have this exposure.

Example of a Laboratory Practical Test Item for Electrical Circuits

The diagram below represents an electric tester.



Use the materials provided to make an electric tester.

1. What happens when the wires X and Y are connected?

   ☐ The bulb lights.

   ☐ The bulb does not light. (Please contact the test administrator.

2. Test the objects provided to see which conduct electricity. Record the name of each object and the results. Place a check in the appropriate column to indicate the result.

Object	Bulb Lit	Bub Did Not Light	Conductor (yes/no)

3. Indicate on the chart which objects are conductors of electricity. Give reasons for your answers.

Disconnect the wires when you are asked to clear your area.

Figure 4.14 Example of a Laboratory Practical Test Item for Electrical Circuits

The preceding illustrations were all selected from the physical sciences. Although assessment items using content from the biological sciences might appear to present additional problems, such as keeping living materials alive, they are none the less important if the test is to reflect the curriculum. We will conclude this chapter with an illustration of an individual assessment item that uses the behavior of mealworms as the content vehicle to gather data on a variety of skills concerned with investigating or problem-solving in science. Mealworms are relatively easy to obtain and keep alive and take up little space. They are often studied in elementary classrooms.

Food Preferences of Mealworms — an individual practical item

The child is introduced to the mealworms and then instructed to offer them food in a way that would help him find out if they preferred one more than the others (Harlen, Black & Johnson, 1981). The necessary equipment, such as food, spoons, stop watch, tray, and ruler, have been placed on the table before the student begins. A student record page is included. Hints are provided for the administrator to give the student if he gets stuck and can not proceed. For example:

> "One way you can try is to make a mark in the middle of the paper. Take some of each food and place it at the same distance from the mark and then put some mealworms on the mark."

If hints are given, the observer makes a record of them. The student page and a list of the skills that the observer focused on are given in Figures 4.15 and 4.16.

Light Through a Jar — a station practical item

The major focus in this item is the planning of a Fair Test by the student (Chiarello, 1989). Later, they conduct the test and write a summary of their investigation. The necessary materials for the investigation are placed on the table in front of the student. This item has been used successfully with fifth grade students. It works best if sunlight is the light used for the activity. Jars for this activity may include olive jars, jelly jars, etc., or plastic vials of various diameters. The student sheet for this practical item is shown in Figure 4.17.

This chapter has illustrated ways to assess problem solving skills in science. We have viewed this as an integrated activity involving planning, doing, analysis, and application of skills. We illustrated the assessment of problem solving with a number of examples of practical tasks.

List of Skills for Mealworm Food Preferences

- Uses hand lens correctly
- Hint given
- Deliberately provides mealworm with choice; i.e., at least 2 foods at once
- Employs an effective strategy such as:
  (i) uses 6 or more mealworms if all 4 food compared at once
  (ii) compares foods in all possible pairs with 1 mealworm
  (iii) tries at least 4 mealworms with one food at a time
- Attempts to provide equal quantities of different foods
- Puts approximately equal quantities of food
- Attempts to release mealworms at equal distance from all foods <u>or</u> arranges mealworms to be randomly distributed around food
- Arranges to release mealworms from points equidistant from foods, <u>or</u> places mealworms randomly around foods
- Arranges for all mealworms to have same time to choose (i.e., puts them all down together or uses a clock)
- Uses clock to time definite events
- Allows about 4-7 minutes for mealworms to make choice (not necessarily timed)
- Examines behavior carefully (to see if food is being eaten)
- Counts mealworms near each pile after a certain time (or notes which food the mealworm is on for strategy (ii) above)
- Makes notes at (a) (however brief)
- Records details such as time of choice and numbers near each food
- Can read stop clock correctly (to nearest second)
- Makes a record of finding at (b) without prompting
- Results at (a) and (b) consistent with evidence (even if only rough)
- Results based on and consistent with quantitative evidence

Figure 4.15. List of Skills for Mealworm Food Preferences

Example of a Student Record Page

Mealworms

Find out if the mealworms prefer
some of these foods
to others. If they do, which ones do
they prefer?

a) Put down here any notes and results as you go along:

b) Write down here what you found about the foods the mealworms prefer:

Figure 4.16. Example of a Student Record Page.

Example of a Student Record Page

Plan and do a fair test to answer the following question:

WHAT EFFECT DOES THE CURVE OF A JAR HAVE
ON THE PATH OF LIGHT? (Use water in the jar.)

For your information:
The jar with the smaller diameter is <u>more</u> curved than the jar with the larger diameter.

1. What is the question you are investigating?

2. What are the variables you need to keep the same?

3. What specific variable will you look at? (What is it that you will observe as you do your fair test?)

NOW DO YOUR FAIR TEST AND THEN RECORD YOUR DATA ON THE BACK. MAKE SURE IT IS CLEAR, NEAT, AND LABELLED.

4. Write a summary of what you found out. (Answer the question you investigated.)

Figure 4.17 Example of a Student Record Page.

CHAPTER 5

METHODS OF COLLECTING INFORMATION:

HOW TO DEVELOP YOUR OWN ASSESSMENT INSTRUMENT

The previous chapters have described and illustrated a variety of ways that have been used to collect information about science achievement and progress in the elementary school. They have also included discussions about the advantages and disadvantages for each of the basic methods and their appropriate uses. The purpose of this chapter is to address the process of *developing and administering* assessment tasks and instruments. It covers planning for balance and program emphasis; written tests; practical tasks; assessment through dialogue; and using student writing and other products for assessment.

Planning for Balance and Emphasis

Any plan for assessment should first take into account the program objectives. Decisions then need to be made about the methods that are appropriate for collecting the information that is needed. For example, if the program emphasizes only the accumulation of facts, then a totally written test with "knowing" (recall) questions might be appropriate. If the program, however, emphasizes hands-on science with application of learned concepts, then the following assessment blueprint might be used in planning:

		written test	practical work	project
knowing	(25%)	25%	—	—
using	(60%)	10%	40%	10%
extending	(15%)	5%	5%	5%

The above blueprint is limited in scope. It might be an appropriate choice for planning assessment at the end of a fourth grade unit about electricity. Lists of the specific concepts and skills that are to be assessed would have to be drawn up. These lists would then be used to construct the assessment items/tasks in the proportions specified by the matrix. In this manner, for example, using the concept of a complete circuit might be assessed by: (a) a paper and pencil item; (b) through practical work; or (c) from evidence gleaned from a final project.

If the purpose for assessing were to find out how well (or how much better) students were able to *perform a test* to determine which materials did, or did not, conduct electricity, then the practical assessment format would be the appropriate choice, (although a practical project might conceivably be appropriate, too.)

For a more comprehensive assessment, such as the end of the term or program, evaluation, the grid in Figure 5.1 might be used (Meng & Doran, 1990). The descriptors on the left-hand side of the grid pertain to the skills appropriate for the particular grade/age in the syllabus used by the school. The columns of the grid list the several formats available to assess these skills. The grid can be used to plan balance and emphasis, as done in the preceding grid, or to list the specific skill in the appropriate cell. All cells might not have the same number of items, with some being left blank. If the purpose were to evaluate a program, not students, individual students would be randomly assigned to only a portion of the total assessment. (Sampling techniques will be discussed in Chapter 6.)

A nationwide study of eleven-year olds in England utilized a variety of formats as part of the program to assess process skills (Harlen, Black & Johnson, 1981). Format choice was based on which was considered most appropriate and effective in gathering accurate data concerning the specific skill(s) (see Figure 5.2). Some skills, such as "using symbolic representation" and "interpretation and application," are assessed by *paper and pencil* tests administered to the whole group. Other skills, including "observation," are assessed during a *group practical*; students working a series of individual tasks with information gathered via written response. A small sample of students was observed while "performing an investigation" (an *individual practical*). The first two methods yield written student responses that can be evaluated after the test is administered. The individual practical, although including a student record sheet, depends largely on teacher observations recorded on a check list.

Written Tests

There is a some cynicism about paper and pencil or written tests for the elementary school youngsters. Throughout this volume we have stressed the need for several forms of assessment. One can certainly agree that using *only* written tests would be unwise. Many important objectives are measured by the group and individual practical test formats. However, it is also true that many important objectives can be assessed by high-quality written tests. Further, high-quality items *can* be written that assess more than just knowing.

There are a variety of item formats that have been used in written tests. Each has strengths as well as weaknesses. Teachers should choose the item format that

Assessment Grid for Elementary School Science				
		ASSESSMENT FORMATS		
	Written	Observation	Group	Individual
	Written			
SKILLS	Test	Discussion	Practical	Practical
Projects	Work			
Knowing:				
Using:				
Extending:				
Planning/ Design:				
Performing:				
Analyzing/ Interpreting:				

Figure 5.1. Assessment Grid for Elementary School Science

best "fits" the objectives and the students being assessed. For example, if the objective is for the student to propose ways of planning a "fair test", perhaps an essay question is most appropriate. If the goal is to see if students can recognize examples of plants and animals, perhaps a matching item is most appropriate.

In cases where several item formats seem appropriate, items could be written which assess the same objective via several different formats. Then one could examine the information obtained from each format and choose which provides the "richest" set of data. From some item formats one can determine what percentage of the students supply or select the correct answer and in addition how many students possess specific misconceptions or how many choose an inappropriate method of calculation. One might also obtain from some item formats information about student writing skills, their attitudes and values about science and school, and their interest in various types of science-related jobs.

other hand, essay questions are generally considered to be subjectively scored because much interpretation and decision is required by the scorer. Later we will discuss strategies and recommendations for scoring essays that can reduce the subjectivity to some degree.

Essay Question

The essay question was the first form of written testing to be used. It was an outgrowth of oral questions which were the predominant assessment technique until the 1800's in the United States. As schooling became available to more and more people, the essay question evolved as a way to assess the achievement of the greater numbers of students.

Essay questions are believed to be capable of assessing higher level reasoning skills. They also can provide a measure of writing, organizational, and communication skills. Essay questions are commonly perceived to be easy to write. However, hastily written essay questions can be difficult for students to interpret and nearly impossible for teachers to score.

Cartoon strips pointedly describe the student frustration with ill-prepared essay questions. Questions, such as "Describe plants" or "Tell me all you about rocks" leave students totally perplexed and understandably so.

Teachers need to limit and focus questions to help students realize what is expected. Otherwise, the students are required to become "teacher mind readers" to determine what the teacher is requesting. For instance, "Describe how plants contribute to the oxygen cycle" provides a much better focus for students. Similarly, "What are some reasons why fossils are found only in certain kinds of rocks."

A common complaint about essay questions is the difficulty associated with their grading. The grading difficulties are described as being of both a quantity and a criterion base. The quantity aspect is based on the sheer numbers of student papers to grade. Unless clear criteria are established for grading, *prior* to the actual scoring, subjectivity can become a severe problem. Teachers may also tend to unconsciously "read into" the answer of "good" students, being convinced that they really knew what was going on, even though the answer did not indicate that.

The major recommendation for the use of essay questions is oriented toward improving the scoring guides. The form of the guide will vary with the nature of the question and the level of the students. Scoring will be discussed further in Chapter 6.

If several essay questions are used in one testing session, teachers should clarify the time to be spent on, or points allocated to, each individual question. Some students would assume that each of five essay questions in a test would have the same "value" as the others. Other students might assume that some are worth more points because they are longer. Such assumptions should not have to be made. Even if all questions are to receive the same number of points, specific directions to that end should be included in the instructions.

Generally, it is recommended that "students should *not* be allowed to choose among a set of essay questions." The major reason for this recommendation is that each question is *not* identical in terms of complexity or difficulty, no matter how hard we may try. Each set of questions that an individual student might select is unique and therefore, the students are in essence taking different tests. Every question used in an examination should be important and therefore should be answered by every student. If students know that they will have a choice of test items, they may choose to study only a portion of the material and "play the odds" on being asked questions covering the material studied. Most teachers agree that students should study all parts of the course, so correspondingly they should be required to respond to all parts of the test. It is possible that students of differing abilities may respond to different items, thereby creating bias within the test. Teachers may also react more favorably to the choice of some questions over others, further clouding the validity of the test.

The last, but probably the most important suggestion is to allow sufficient time for preparing the question. One should write the first draft of the question, obtain feedback and comments from colleagues, and then rewrite the question. At this point the teacher should prepare the scoring guide and get feedback from colleagues before subsequent revision of the question and scoring guide. These steps need to be followed before trying the item with students. Once student responses are available, additional revisions may be necessary. Essay questions are an important tool for the assessment of elementary school science programs. As with all instructional and assessment materials, care and skill are needed to produce a high-quality instrument.

Completion Questions

"Completion questions" also require the student to supply an answer, but all that is usually expected are a few words or phrases. An advantage of this item format is that students must provide the answer—they are not able to select or choose from options available. These items are quite easy to construct and quick to score. The major disadvantages of completion items are the difficulty in assessing objectives above the recall level and the problem with controlling the range of the acceptable answers.

In the large majority of cases, students will easily recognize the text or instructional material from which the question came and provide the expected answer. There are other cases, when students respond with specific examples instead of the desired more general term. These cases require the teacher to make a choice, often in the midst of grading the papers. Other problems are completed by "creative" answers, such as the following:

Matter occurs in three states: ________________________,

________________________, and ________________________.

This item has been answered in many ways, e.g., New York, New Jersey, and Pennsylvania. Although patently absurd, this answer is a correct response to the item as it is written. This item could have been better worded in the following way:

Based on temperature and pressure, matter may exist in each of these three phases or states: ______________________________, ____________________, and ________________________.

Suggestions for writing and revising completion items will be summarized below. The caution to "avoid vague questions" was illustrated by the question above. Likely, it is crystal clear to the teachers who wrote the question what was expected. However, what is more important is that the students can correctly interpret the task. Often having a spouse, a friend, or a colleague read a test question will uncover ambiguities that would confuse students. Usually completion items with one or two blanks will work fine. It would appear that the more blanks that are provided, the more information has been gathered. While that is true, statements with excessive blanks become almost uninterpretable because so much of the substance of the sentence is deleted. The following is an example of such a multi-mutilated item:

Most <u>green</u> plants produce <u>sugar</u> from <u>water</u> and <u>carbon dioxide</u>.

In this case, eliminating any two of the blanks will improve the item.

It is usually easier for students to respond when the blanks are at or near the end of the statement. With just one reading of a sentence, a student can determine the meaning of the statement and will provide the missing information, if known. For example:

POOR: ________________________ are the hair-like structures by means of which paramecia move.

BETTER: Paramecia move by means of hair-like structures called ________________________.

The second item can be completed more readily by students because of its more direct approach. When they reach the blank, they should have all the information they need to write the answer, if they know it.

Avoid extraneous clues to the correct answer. Sometimes clues to the answer are unintentionally provided by the grammatical structure of the item. For example:

A reaction among the subatomic particles is what scientists call a _____ reaction.

The use of "a" in this item would indicate to the alert pupil that "atomic" cannot be the answer because "a atomic reaction" would be incorrect English. This item could easily be improved either by using the "a/an" phrase or by changing the form of the nouns from singular to plural:

> Reactions among subatomic particles are what scientists
> call _______ reactions.

Another common extraneous clue is given by using short blanks for short word answers and long blanks for long word answers. The same length of blank should always be used to avoid cuing to the students the relative length of the word or phrase desired. A similar mistake is to indicate by the number of blanks the number of words in the correct answer:

> The gas that makes a cake rise is _______________ _______________.

Realizing that the correct answer has a compound name, students will probably not answer with the names of other likely, single-word gases like "oxygen" or "nitrogen."

Matching Questions

The "matching" item format is favored by many teachers because it requires relatively little reading by the student and covers content very efficiently. The scoring of these questions is relatively simple and objective. The guessing or chance factor is minimal if the following guidelines are followed. The major drawback to completion items is that they can assess only a certain range of objectives, mostly at the knowing level. Students are commonly asked to match exemplars with concept names, symbols with names, etc. In other cases, items could require the matching of structure with function, cause with effect, graphs with relationships, etc. Items so constructed might assess levels higher than knowing.

The major guideline for constructing matching items is to "keep the lists homogeneous." The specific words within the list of stimuli and responses should be clearly identified as within some area or domain. The following example demonstrates an item with a very heterogeneous set of stimuli and responses:

Match the following with where they are found:

____	1. Human	A. Solar System
____	2. Bird	B. Nest
____	3. Venus	C. House
____	4. Car	D. Hole
		E. Garage

In this example, the stimuli come from such a wide range of fields, that completion is based only on a superficial knowledge of words, not from understanding science concepts. An example below illustrates an item with a more homogeneous set of stimuli and responses:

Match the following with where they are found:

____	1. Fish	A. Cave
____	2. Snakes	B. Water
____	3. Bear	C. Nest
____	4. Bees	D. Hole
		E. Hutch
		F. Hive

Whenever the procedure is changed, the student must be informed. For instance, in some matching items a teacher may wish the student to match each response with several stimuli, while in other items several responses may be matched to a simple stimulus. Specific directions and sample items should be used if this is the student's first exposure to this kind of matching procedure.

These lists should be kept relatively short, especially for younger children (no more than five or six elements). For many students, choosing their responses to matching questions involves several re-readings and comparison of a list of stimuli and responses. They try to make logical matches, given the basis or criteria for the question. If the lists get longer, the searching process creates a greater number of possible comparisons to consider. As was just implied, one should state "clearly the basis for the matching exercises." In many cases it is quite obvious what is being expected. However, a short statement is a simple way to eliminate potential confusion. Another suggestion to simplify the task for the students is to arrange the stimuli and responses in some appropriate order. In many cases that may be simply to list the objects or items in alphabetical order. If the responses are numerical, they should be listed in some order, either descending or ascending. Other logical ways of ordering matching lists will be obvious to the teacher involved.

One should always provide more responses than there are stimuli. Usually two or three more will be adequate. If one does not follow this guideline, the last few matches require little thinking. Students get credit for each response, even if they merely match the remaining response with the remaining stimuli. These lists of stimuli and responses should be printed on the same page. This suggestion is based solely on facilitating student reading and responding.

True/False Items

The advantages of true/false questions are several: they are easy to write, they cover a lot of content, and they can be objectively scored. However, the problems cited with these items are many. True/false items primarily assess knowing level objectives. They are very difficult to qualify simply, and the guessing factor is large. Most true/false statements are lifted verbatim from the text or from prepared instructional material. When that is the case, it is clear that recall of specific information is being assessed.

The following example demonstrates the difficulty of adequately "qualifying" statements:

T F Water boils at 212°.

The better student may mark this item false, because the temperature scale is not indicated, the atmospheric pressure is not mentioned, or because the statement is not true for sea water.

While most true/false items appear to be very simple, they are often open to many interpretations and questions:

T F The Earth is nearest the sun in December.

The intended answer is "true" on the basis that during December the Earth is in the position in its orbit which is closest to the sun. However, a student could interpret the statement to mean that in December the planet nearest the sun is the Earth, an interpretation that yields a "false" answer.

A major complaint against true/false items is the larger guessing factor. If students "blindly" guess, they earn scores around 50% correct (assuming there are the same number of items keyed true and false). Further, if they know the correct answers to a few statements, their scores are likely to be above 50% but their "real understanding" is at a much lower level.

There are a number of suggestions for writing true/false test questions. The main recommendation is to write statements as clearly as possible. Qualitative words, such as "many," "better," "long," or "heavy," leave many varying interpretations. Statements are much more clearly understood when precise descriptors are used: 12 meters, 16 kilograms, etc.

Teachers should avoid "trick questions" and eliminate "window dressing." If "trick" phrases are used, the test becomes a measure of test wiseness and not an assessment of science. "Window dressing" is a label for extra words and phrases. In an attempt to make the item more practical, personal, relevant, and appealing, we often add extra statements and descriptors. Most of the time such attempts merely add reading demands and confuse students.

Another problem is double-barreled items that are partially true and partially false. These confusing items can obscure the important point that is intended to be assessed. Some people believe that an item can be made more comprehensive by incorporating a number of ideas:

T F Because evaporation is a cooling process, a swimmer feels chilled when coming out of the water on a windy day.

Instead, it generally clouds the purpose of the measurement. Students may feel that they have grasped the central idea by reading one of the parts of the item, overlooking the important part of it which is hidden or obscure. We should place the important words or phrases in a prominent position in the sentence, so students can easily focus on the important elements and choose their responses.

One should include approximately the same number of true and false statements within each test. Teachers often "naturally" write more true than false statements. Many statements can be easily modified to accomplish a better balance of true and false statements.

If sentences must include "negative" words, such as "not," "never," and "except," these words should be highlighted by capitalization and/or underlining. This will minimize the chance of a student quickly reading the statement for general understanding and Missing that one small word that would change the entire meaning of the sentence.

It is recommended that students indicate their response by circling a "T" or an "F" printed on the test booklet or answer sheet instead of having the student print a "T" or an "F." Creative students have been known to argue that the symbol they wrote (that we thought was a "T") really was intended to be an "F."

Teachers have evolved modifications of true/false items in attempts to minimize some of the problems with true/false test questions. One such modification is to require that students "correct" statements that are false — in such a way as to make them true.

In these items students are directed to focus on one key word or phrase (which is normally underlined or bracketed) and to use this element as a basis for deciding whether the item is true or false. The following example with directions alerts students to the features unique to the modifications:

DIRECTIONS: In each of the following true/false statements, the crucial element is underlined. If the statement is true, circle the "T" to the left. If the statement is false, circle the "F," cross out the underlined word, and write in the blank space the word which must be substituted for the crossed-out word to make the statement true.

T F Taurus 1. The Pleiades and the Hyades are in the constellation Orion.

Other teachers have created "multiple true/false" items. In these items each response must be separately judged as to whether the stem and each option create a true or a false statement. This form of a true/false item is very similar to the multiple choice test item. The following item (Figure 5.3) illustrates this item format (Harlen, Black & Johnson, 1981):

Example of Multiple True/False Items

Some girls had a collection of toy cars of different colors and sizes.
This is the investigation they were doing with them.

1. Beth and Eve stood at one end of the playground and Joan at the other end
2. Beth and Eve put different colored large cars on the ground one at a time.
3. Joan tried to tell the color of each car as it was put down.
4. Then they did the same with the different colored small cars.

What would they find out from this investigation?

Read the sentences below. For each one put a plus (+) if you think it is something they would find out and a zero (0) if it is not.

☐ Which colors could be seen from a distance.

☐ Whether the distance made any difference as to which colors could be seen.

☐ Whether Joan was better at seeing the colors than Beth and Eve.

☐ Whether the size of the car made any difference as to which color could be seen from a distance.

Figure 5.3. Example of Multiple True/False Items.

Multiple Choice

This form of test item is used widely in the assessment of instruction in U.S. schools. By the time most students are in the second grade, they have experienced some sort of multiple choice test item. In the early grades the form of the item is such that the teacher or test administrator reads the question aloud to the class, and each student chooses among the several pictures that represent the possible answers.

For many teachers the multiple choice item has fewer problems and more benefits than the other item formats. The conclusion should not be interpreted to mean that the item format is without fault or that it should be used to the exclusion of the other forms of test questions. Some of the benefits cited for multiple choice questions are the limited guessing factor, broad coverage of content, and objective scoring.

Some problems can be expressed more simply and clearly in the incomplete statement style. On the other hand, some situations seem to require direct question stems for most effective expression. Thus, neither of the following items can be presented quite as neatly in the other forms:

Which is a mammal?

1. rat
2. mosquito
3. rattlesnake
4. spider

Flowers are necessary to produce

1. bulbs
2. roots
3. seeds
4. cuttings

Teachers should use the style that seems most appropriate for the particular objective at hand. If in a given instance it seems to make little or no difference which item type is used, the teacher should choose the style that can generally be handled most effectively by the students. There is no evidence that either type is inherently superior to the other.

Teachers who have not had experience in writing multiple choice items may find that in the beginning they will tend to produce fewer technically weak items when they try to use direct questions than when they use the incomplete statement approach. The reason is that in the incomplete statement approach it is often difficult to arrange qualifying phrases or words to produce a perfectly clear statement. In addition, because of its specificity, the direct question induces the item writer to produce more specific responses.

Many of the recommendations cited for writing the other forms of test questions are also applicable for multiple choice items. In addition, we will summarize some suggestions specific to multiple choice test questions.

Each of the options should be plausible and attractive to at least a few of the students being tested. If the question requires the student to select a winner of the Nobel prize in science, little is gained by including names of rock stars or athletes as distractors. Test-wise students can spot weak distractors and choose "correct" responses without possessing the knowledge intended to be measured. If implausible distractors ("deadwood") are included, the question is no longer useful for what it was designed to measure. A four-choice item becomes essentially a two-choice item if two of the options are not selected by any students.

Options that are added to contribute humor to the testing situation should be recognized as accomplishing that and only that. In reality an item with three realistic choices and one "humor choice" is in effect really a three option item. A good scurce of statements to use with the multiple choice items are the responses which students have offered to a question of that content, organized in an open format. Options should be written so as to represent known or likely misconceptions or errors for students at the level of sophistication. The "closer" an option is to the correct choice, the higher the lev l of discrimination which is being assessed.

All responses should be independent and mutually exclusive. Each response is related to the other in the sense that all are potential responses to the stem, but they must not overlap or cover a common range of possibilities. Some students will be able to spot these inconsistencies and will then, in effect, be responding to a different item. The following set of options is an example of this problem:

1.less than 20%	3. more than 40%
2.less than 40%	4. more than 60%

Many students will spot the internal problem: if "1" is correct, "2" is also correct; if "4" is correct, so is "3." Since "2" and "3" cover the spectrum, options "1" and "4" will be ignored, making it a two-choice item. The options for this item could rather be:

1.less than 20%	3. between 40% and 60%
2.between 20% and 40%	4. more than 60%

A common problem in constructing multiple choice options is making the correct choice shorter or longer and more complex than the distractors. This can happen unconsciously as the instructor writes the item. Students can also get clues from grammatical inconsistencies, specific determiners, and key words. All responses must be grammatically consistent with the stem and parallel with one another in form: e.g., all action statements or all people.

The responses "none of the above" and "all of the above" should be avoided. There are situations in which these responses are highly appropriate and should be used, but more often they are added when more relevant distractors are not easily available. These options particularly lend themselves to being selected for the wrong reasons or for no reason at all. Thus, in order to maintain their validity, they should constitute the correct answer only on occasion — but certainly no more often than the other options are "correct" (approximately l/k of the time, where "k" is the number of options per item). If items are not carefully written, students may read particular meanings into the items and options and argue that "none of the above" is a defensible answer. If constructing viable distractors is a problem, another item format should be considered before using a multiple choice item with several weak distractors.

For items in which the choices involve numbers for calculations, a system for constructing item choices will be discussed. If the correct mathematical calculation for solving a problem is subtraction, the other choices could be developed by using other elementary math operations, such as adding, multiplying, and dividing (which it could be, too). Another technique for constructing options is to use some of the values cited in the stem as possible answers. Each of these procedures is better than simply choosing numbers randomly above or below the "correct" choice.

The choices should be arranged in some order to facilitate student reading and comparison. One such order is an alphabetical list of words or phrases, descending or ascending order for numerical answers, and grouping of responses that have some element in common (e.g., "increasing" or "decreasing"). The following exception to this recommendation is noted for responses which include numbers between 1 and 4. Students may confuse the absolute value of the answer with the keyed number of the response, as:

POOR	1. 1	3. 4
	2. 3	4. 6
BETTER	1. 1	3. 3
	2. 6	4. 4

Each item in a test should be independent of all other items. The content or wording of one item should not give away the answer to other items. It is not recommended to make the successful completion of an item depend upon the answer to the previous item. Such sequential items may serve one kind of purpose, but for the inexperienced teacher and student they will often create problems.

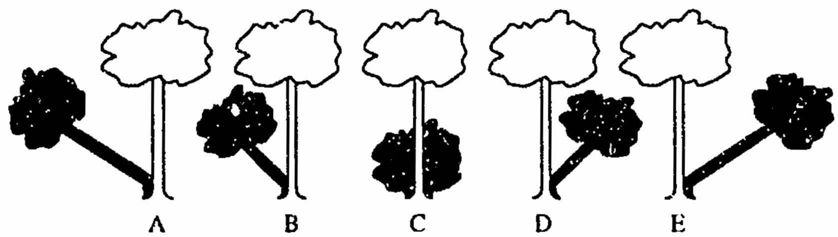
Avoid patterns in the order of correct choices within a test. A visual scan of a keyed answer sheet can often point up such common patterns as an "A-B-A-B" or an "A-B-C-D" sequence. Similarly, the position of the correct response among the available choices should vary; e.g., response "A" should be keyed correct approximately as often as responses "B," "C," and "D," and so on. Counting the number of times each response position is keyed "correct" will determine existing proportions and point up imbalances. If necessary, the responses in some items can be switched to achieve a better overall balance.

Pictures, diagrams, or tables of data are excellent ways to clarify the information presented. These techniques, when appropriate, will also minimize the ever-present problem of excessive reading demands. Unless the identical information and context that were used in the instruction are present in the item, it is probable that students will be required to interpret and translate information to determine their answers. Another value of such techniques is that they can simply be changed for later use by inserting new information or numbers.

The first test item that follows, with diagrams of trees and their shadows (figure 5.4), is from the tests administered as part of the Second International Science Study (SISS, 1989). The next two items, utilizing both a table of data and a visual portrayal of thermometer readings, are also from SISS (Figure 5.5).

Example of a Test Item Using Pictorial Information

At different times during a sunny day a tree was seen to have cast shadows of different lengths as shown in the diagrams below. Which diagram shows the shadow at mid-day (12 noon) ?



A. diagram A
B. disgram B
C. diagram C
D. diagram D
E. diagram E

Figure 5.4. Example of a Test Item Using Pictorial Information.

Example of an Item Using Information in Tables

The next two questions refer to the following table which shows some temperature readings made at different times on three days.

	6 a.m.	9 a.m.	12 noon	3 p.m.	6 p.m.
Monday	15°C	17°C	20°C	21°C	19°C
Tuesday	15°C	15°C	15°C	10°C	9°C
Wednesday	8°C	10°C	14°C	14°C	13°C

1) When was the highest temperature recorded?
   A. Noon on Monday
   B. 3 p.m. on Monday
   C. Noon on Tuesday
   D. Noon on Wednesday
   E. 6 p.m. on Wednesday

2) Which of the following shows the temperature at 6 a.m. on Wednesday?

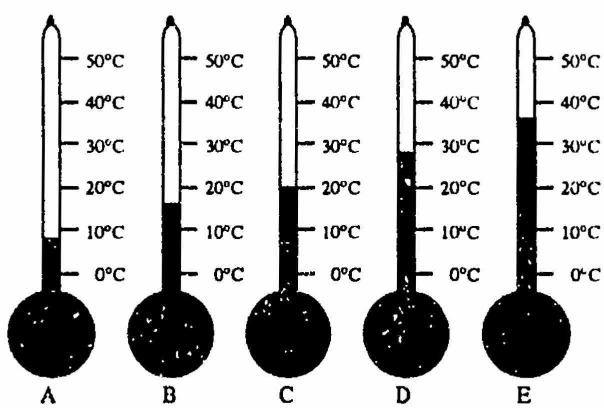


Figure 5.5. Example of an Item Using Information in Tables.

Writing assessment items of any format requires a number of competencies and skills. The test writer must thoroughly understand the content being assessed, both from the subject matter and from the instructional viewpoint. The writer needs to be very familiar with the usual activities and material through which students normally experience the content and processes being assessed. One also needs to be aware of the cognitive development levels normally associated with students at the particular age or grade level, as well as the reading and mathematical skills.

Given the above, the teacher still needs considerable practice to become a developer of high-quality test questions. As part of the practice, it is very helpful to obtain comments and feedback from colleagues and students. Lastly, one should realize that, as is true with other teaching skills, time and experience are needed to develop high levels of efficiency.

As the task of writing high-quality test items is time-consuming, many teachers keep the items for use at a later testing time. Many teachers also collect items from curriculum projects, research studies, and colleagues. These collections are often called "item pools" or "item banks." Some teachers store these on index cards, while others place them in files within a microcomputer. It is most helpful to categorize each item as to content topic, objective level (knowing, using, extending), and percentage of students that answer the item correctly. Items can be selected from this pool or bank for a given test administration. If the pool size is large enough, any one test will have minimum overlap with adjoining administration.

Practical Assessment

Practical exams have been around for quite some time but are still a rarely used form of assessment in the American elementary school. The term "practical" denotes the use of manipulative materials and the interaction of the students with them and the phenomena being studied. "Lab practicals" have been used in secondary schools for the purpose of assessing skills involved in laboratory work, but minimally at the elementary school level. This omission is of no great consequence *if* the focus of science education is only on the accumulation of a body of knowledge. It is an extremely serious oversight, however, if the program objectives include investigation and problem-solving, as most current day programs do. Only through the child interacting with materials and/or the phenomena can we determine if many of the objectives of such programs have been met. Practical exams are not usually used for assessing information learned ("knowing"). Their strength, and purpose, lies in the potential for the student to demonstrate ability in analyzing, formulating logical answers based on observations, measuring, using equipment, and devising approaches to solve problems. They generally require the student to use the higher cognitive abilities, although this will vary according to the exam. Practical tests can differ in what they test. In general, they are most effective in testing the process objectives of a program.

There is a variety of forms that practical assessment can take. Some may resemble the "lab practical" test, also called "station" format. Others might involve the observation of a student(s) during normal activity or a situation specially constructed as a "test." There is also a variety of adaptations that can be made to suit special situations and purposes, such as combining practical assessment formats with other formats; e.g., a written portion or dialogue between teacher and student(s).

Group Practicals

As the title implies, a *group* of students can be assessed by this method. While it is possible to simultaneously assess all the students on the same task, the need for numerous sets of duplicate materials often makes this method impractical. Sometimes a demonstration or audiovisual presentation to a group may be the stimulus for a series of questions which are answered by individual students. More often, a station format is used; children rotate through a set of different tasks. As the station format is an effective method, planning and set-up will be described in some detail.

The Tasks

After a balanced assessment plan has been constructed, activities are selected that could elicit the behaviors that we wish to assess. Rarely is it appropriate to use the same instructional task as an assessment item. Instead, a similar activity is chosen with the same concepts and skills. Although it may be tempting to include many of the skills and concepts to be assessed in one ite. 'his can make the results difficult to interpret. Therefore, it is often best to limit the number of skills and concepts assessed in each item. For *process* assessment this can be accomplished by choosing an activity that is familiar to students and therefore not dependent on "school" science, such as observing similarities and differences of toy animals as illustrated in Figure 5.6 (SIS ,1986). It is also possible to give the students the science information that might interfere with process assessment.

There are other factors to consider when choosing or developing activities for assessment items. Some that we have found to be important are given below:

- One segment of a task should *not* be dependent on thesuccessful completion of a previous segment unless it is to be used for comparison and to determine a relationship.
- Items can vary in how much structure, or help, is given to the child in organizing a response; but anything *not* given should be considered another variable being tested, thus adding complexity to the task. Try to keep the task simple for young children.
- Do *not* choose an activity that needs a lot of explanation or directions.

Example of a Group Practical Item

Two plastic animals are on display before you. Look at them carefully.

1. Other than color, list three ways in which they are alike.
   1. ______________________
   2. ______________________
   3. ______________________

2. Other than color, list three ways in which they are different.
   1. ______________________
   2. ______________________
   3. ______________________

Figure 5.6. Example of a Group Practical Item.

What activities to use for any particular station needs some forethought, for not all tasks take the same duration of time to complete. For example, reading a pre-set scale takes less time than the actual weighing or measuring of a precise quantity, which often takes longer than we anticipate. Time also has to be allowed for assembling and dismantling equipment, if that is required, and for reading and understanding instructions. What type of record and how much a student is expected to record will affect the time needed to complete an activity. Although the test needs to be timed fo: smooth flow, adequate time must be allowed for the more deliberate child to complete it according to his/her ability and skill.

It is recommended that you try the tasks yourself in order to determine the relative time necessary to complete each task. Revisions can then be made, or related "short" activities can be grouped at some stations to equalize the time interval. Elementary school children are usually quite good-natured about helping us try things out if they are told before hand. A trial test, run in this manner, can be very helpful in debugging the process.

The Student Booklet

The written instructions and questions must be constructed in keeping with the recommendations for clarity and simplicity discussed in the preceding section on written tests. A carefully constructed student record sheet or booklet is usually the preferred method for presenting the information and gathering data from the students. Each page should include: the question or task, even if given orally; places for the student to respond; and any other appropriate information, such as what to do when the task is finished. Figure 5.7 illustrates an item used for an international assessment of fifth graders (SISS, 1986).

Example of a Student Test Booklet

Before you is a "Testing Sheet," a Q-tip, and some cooking oil. Dip the Q-tip in the cooking oil and then rub it against the "Testing Sheet" in the space labelled "cooking oil" and observe what happens.

1. Describe what you saw happen when you rubbed the oily Q-tip against the paper

____

____

In a container you will find different plant seeds. Determine which seeds contain oil.

2. Describe the procedure you used to determine which seed(s) contain oil

____

____

3. List the name(s) of the seed(s) which, according to your test, contain oil.

____

____

TESTING SHEET FOR COOKING OIL

Seed 1 ________ Seed 2 ________

Seed 3 ________ Seed 4 ________

Seed 5 ________ Seed 6 ________

Figure 5.7 Example of a Student Test Booklet

Student booklets need careful planning, trial testing, and revising. We have found the recommendations below to be helpful in preparing them.

- Whether the question has been given orally or not, it should be repeated in the booklet.

- Diagrams are helpful and should match the actual materials being used by the student.

- Adequate, lined space should be provided right after the question for student data and responses.

- Minimize the dependence on reading and writing ability. Scientific terms or any unfamiliar words should not keep the student from being able to complete

  the task. Children with reading problems may need help in word recognition.

- Use large, clear print and uncluttered pages. Sections should be well defined.

- It is best if all of the information (questions, directions, student responses) can fit on one page or continued on a facing page.

- Expected student response should be delineated for the child. If comparisons are to be made in terms of similarities and differences, state this. If a specific number of observations is required, state the number. If interpretation or identification of a pattern is expected, ask such a specific question.

- Each page should be clearly marked with the identification number or letter of the corresponding station.

- Include a place for the student's name and make sure it is filled in *before* starting the test.

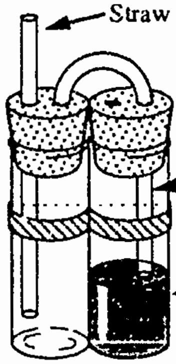
- Read general instructions for the test before beginning. Have a copy of this on the front page of the booklet.

- Make sure the test booklet is securely stapled together.

Figure 5.8 illustrates a student page in which a diagram of the actual materials involved is used to convey necessary information with a minimum of reading required (SISS, 1986)

Example of a Student Test Booklet with a Diagram

Before you are two containers. One contains a blue liquid. Place a straw in the stopper of the container that has no liquid in it and blow into the straw for about one minute.



1. What change has taken place?

____

____

____

____

2. What is an explanation for this change?

____

____

____

____

Figure 5.8. Example of a Student Test Booklet with a Diagram.

The student responses in the booklet provide a record which can be examined at a later time. A scoring guide for each item needs to be developed based on the behaviors which the activity intended to elicit. This is done before the test is used. (See Chapter 6 for the development of scoring guides.)

Materials and Equipment

When developing practical assessment items, major concerns involve obtaining appropriate and sufficient materials and then managing and organizing all these materials before, during, and after the test. Although this might seem to be a tremendous problem, a little planning and practice will make the process surprising easy.

A detailed list for each task should be developed to guide the acquisition and use of equipment and materials The list in Figure 5.9 is for a task which is part of the SISS process test (SISS, 1986).

An Equipment and Materials List for a Lab Exercise			
Exper-iment	Quan-tity	Item	Comments
1	3	Styrofoam cups (7 oz.) with lip about 3/4 way up	Each cup should contain water at about 0°C. (Water could be chilled with ice and stored in large styro-foam cup with lid.) Label small cup "X."
	3	Styrofoam cups (same as above)	Each cup should contain water at about 50°C (hot tap water). This water should be stored in a large cup with lid. Label small cup "Y." *Please note that the hot and cold water should be distributed just be-fore the starting time.*
	2	Thermos jugs (at least 25 oz. and with tight covers)	For storing the hot and cold water.
	2	Large styrofoam cups (16 oz.)	For mixture of hot and cold water.
	2	Plastic spoons	
	[illegible]	Thermometers 0°C to 100°C	With only one temperature scale. Thermometers should be glued to the metal back.

Figure 5.9. An Equipment and Materials List for a Lab Exercise.

The list appears to be straight forward and simple, as is intended. However, much trial testing and revision of equipment occurred before the list was finalized. As you may be developing your own practical assessment instruments, be sure to prepare such a list for your future reference and possible use by others. Most importantly, be as precise and detailed as possible. For instance, you must remember that not all styrofoam cups are constructed the same way, nor are all "clear" cups of the same transparency or shape. All insulated wires with clips are not able to withstand repeated use by students. Plan to include extra materials in case something breaks or spills. Include a number of waste baskets for dumping waste.

Once the materials and equipment have been procured, it is time to assemble it for each station. We store the materials for each task in a labelled zip-lock plastic bag or in a box if more room is needed. One can easily locate the appropriate bag or box and assemble the station quickly. The materials can be replaced and stored for future use.

Before the students arrive for testing, one needs to check each setup to be sure it is working correctly. This phase is aptly called "trouble shooting." A number of unanticipated things can occur. For example, we found that some public water sources were acidic enough to turn Bromothymol Blue solution yellow before the student exhaled into it. A bulb in an electric circuit can also fail to light due to a defective part (bulb, dead battery). One also needs to check the supply of ice cubes and hot water, as they are commonly required to be replenished during the assessment. Other specific checks will be dictated by the particular tasks which are developed.

When we try an idea for a practical assessment task, we occasionally have a "flop." This is often due to the materials that we have selected for the student to use. Thermometers slip on the frame or have two scales on them, which confuses the students and exacerbates our attempts at scoring. Materials can be too fragile for repeated use or too complicated for reassembly in a hurry. They can also be unfamiliar to the children and thus add another unintended variable to the test situation. But we learn through practice. What we have learned we would like to pass on to you:

- Each station should be clearly marked with a large number or letter corresponding to a similarly marked page in the student booklet.
- The stations should be visually divided or separated so that the materials do not become mixed.
- All materials necessary to complete the task should be included at each station. Children should *not* have to fetch things or go to another part of the room to use a piece of equipment.
- Use simple, sturdy materials that are familiar to the students.
- Keep stock bottles and other expendable materials in a separate place to avoid accidental contamination.
- Provide marked waste containers for materials that will not be reused.
- Have extra materials to allow for "accidents."
- The actual materials being used should match the diagrams in the student booklet.
- Plan life science tasks ahead of time. Things need time to grow. Have extras in case some do not survive.

Organization

The organization for the group practical takes some careful attention to detail in order to ensure a smooth operation. It becomes much easier with planning and practice. Stations are usually set up around the room to which the students rotate in order. For example, see Figure 5.10.

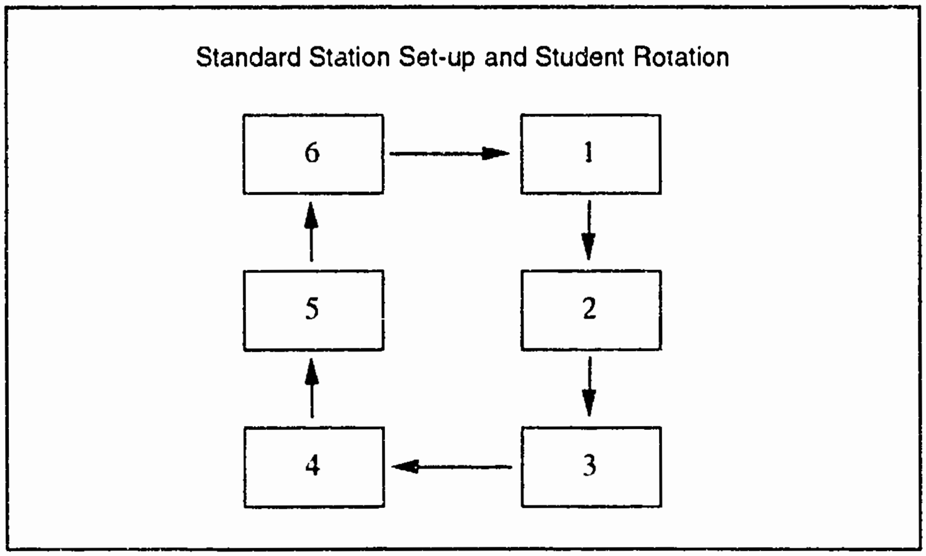


Figure 5.10. Standard Station Set-up and Student Rotation.

In this example each of six students starts at a different station, then proceeds around the circuit, and ends at the station preceding the one where he/she started. A time limit is given for completing the tasks at each station (e.g., 8 minutes). Students are given a one-minute warning before being told to move to the next station. If they finish early, they are told to wait at their station.

Duplicate setups can be made to allow for more students to be tested simultaneously; or a setup can be used in conjunction with a written segment of the test which occupies the rest of the class during this time. If combined with a written test (no materials to interact with), two or three stations might be a reasonable number to expect of fourth through sixth graders. More than six stations is probably too hefty a load for elementary school students, though it depends on what is expected of the students at each station. Although most elementary school children are not accustomed to this sort of procedure, experience shows that they adapt quite readily to it and even enjoy the activities.

We have found the following arrangement of stations to work well (Figure 5.11). It was used for the SISS process testing in the spring of 1986 and accommodates the testing of twelve students per session.

Students sit on the inside of the tables in the diagram and move two stations at the completion of each task. For example, a student starting at the A2 station in the top half of the diagram would move to station A3 after the eight-minute interval and then on to A1 in the same upper half of the diagram. As one can see from the diagram, a student never sits next to a station that he will do next. Further, the students are facing outward since they sit on the inside of the circle, so they would have to work hard to oversee someone else.

For the particular test situation described above, two adults were used. This was because several stations needed materials replenished. The supplies were

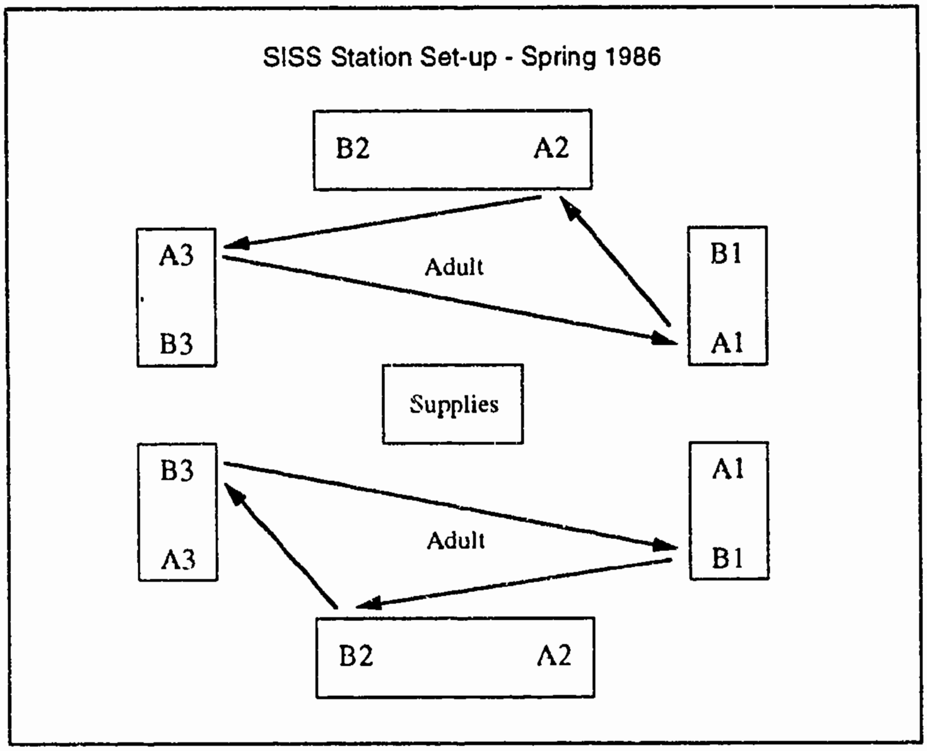


Figure 5.11. SISS Station Set-up - Spring 1986.

placed on a table between the two adults. The people administering the tests were trained but were not necessarily experienced teachers.

Role of the Administrator

During group testing the administrator is kept busy. The major tasks are to: (l) give directions to the students, (2) replenish materials, and (3) monitor student performance.

Before starting the test, the adult in charge explains the procedure to the students. How they are expected to care for the materials at the stations is also explained; e.g., do *not* take the materials apart unless specifically instructed to do so, clean up after using them and put them back the way they should be for the next child, and ask the adult to check if the materials do not look "right" before starting the task. The students are also told what behavior is expected of them and what they should do if they finish a station early, such as restore the station for the next student, read a book (only if they have it with them), do not talk or leave their seats, etc. Even though the instructions may be written on a covering page, they should be read aloud for the students, and then any questions about procedure answered before the test begins.

During the test the administrator manages the materials. This includes monitoring the performance to detect equipment or materials problems, as well as replenishing the stations. If a station needs to be replenished for the next student, the administrator might ask the student at that station if he/she is finished *before* the time is up and replenish *then* in order to save time during the change of stations. Otherwise, it is done during the change. If no other adult is available to help, especially when first trying this method, it may be possible to borrow children from another class to help restore each station for the next child. There are usually some stations that need additional attention. These should normally be identified ahead of time.

There is usually little need for the administrator to interact with the students once the test begins. However, the students will need to be told when to move to a new station and to be given a one-minute warning preceding the move. Some children may need occasional help with word recognition or clarification of procedure. As with any test, the adult must be careful when responding to student questions so as not to give away the answer to a question on the test. Students should be monitored to make sure they are at the right station and in the correct place in the booklet.

Adaptations

Adaptations can be made to the method described above to suit special conditions and needs. Film segments, slides, teacher demonstrations, and pictures may be used when appropriate. However, these adaptations are not the same as having the children manipulate the materials themselves or study the phenomena directly. For example, what is shown in a film segment or slide is pre-selected and often cues a child to the relevant variables.

With any of these adaptations more children could be assessed on the same task at the same time. The instructions could then be given orally. This might be helpful in assessing the progress of young children or of those with limited skill in reading.

Other modifications are possible. Doubtlessly, you will think of other changes as you use this method of assessment.

Individual Practical Assessment

The evaluation of student skills in performing investigations and solving problems can be accomplished by means of individual practical assessment. This method involves one adult observing one student as he/she goes about solving a problem or answering a practical question. Although fewer students can be assessed than if a group practical or paper and pencil format is used, more data can be gathered about one child. Research seems to indicate that the information gathered in this manner is a much more accurate reflection of a student's real ability to plan and perform an investigation (solve problems) than that which is gathered by paper and pencil tests. Because of this, the method has attracted considerable attention recently, especially in Britain (Harlen, Black & Johnson, 1981) and the United States (NAEP, 1987). It promises to be an extremely useful tool for student and program assessment in elementary school science.

The Tasks

The activities that are used for an individual practical are, in many ways, similar to those chosen for the group practical. They resemble the learning task and require the use of the same skills and concepts. However, they are usually more extensive and take longer to complete than those chosen for a group practical task. The number of skills and concepts, therefore, are not as limited. Most often, the activity involves a problem or question to which the student attempts to find an answer through the manipulation of materials. For example, "What foods do mealworms prefer?" or "Does the ball that bounces the highest on one surface, bounce the highest on all surfaces?"

The most effective tasks pose a real problem of interest to the student. ("Which type of paper towel is the best would not interest most fourth or fifth graders.) The tasks need to be chosen or developed with the same care and considerations given to group practical tasks. Some topics that have been used for assessing process skills in the elementary school are:

- dissolving sugar lumps and loose sugar
- investigating the food preferences of mealworms
- investigating the distance a windup toy moves relative to the number of turns wound up
- how differences in strings affect the sound made with each
- determining which boat shape can sail the fastest
- lighting a bulb in a defective circuit
- investigating balls bouncing on different surfaces
- mixing water of different temperatures

There is, obviously, an interaction between the process skills and the background set of concepts that the student has at his command in such situations. Therefore, if the primary purpose of assessment is to determine how well the student can conduct a "fair" test, the content for the task might be something not dependent on "taught" science. Bouncing balls might be an appropriate choice for this purpose, as it could be assumed that it is a common childhood experience. If the primary purpose of the assessment is to determine if a student can apply a concept to a new situation, an appropriate task might be to ask the student to make the bulb light in a faulty circuit. This latter task could give an indication of how well the student is able to apply the concept of a complete circuit to a problem situation.

Assessment of the student's skill in using simple laboratory equipment, such as a thermometer or balance scale, are often done in the context of doing an investigation. A child who is investigating the mixing of two quantities of water of different temperatures would indicate to the observer whether he was able to use a thermometer to determine the correct temperature. The examiner might then intercede and instruct the student on the proper use of the thermometer, so that the student would not be kept from finishing the activity. This would be noted in the examiner's record. Skill in doing a simple test (e.g., whether a material contains starch) could be used for an individual practical. Information about how starch reacts with iodine would, or would not, be given according to your purpose. (However, as with group practicals, lack of success on one section of the activity should not keep the student from completing a subsequent section.)

Role of the Administrator

While the supervisor of a group practical might need some training in how to administer the test, the person does not need to be someone knowledgeable about the activity to be used. The person supervising an individual practical, however, does need to be knowledgeable about what to observe. Judgments about how well a student is able to solve a problem in a scientific manner are made primarily on the basis of the records of the examiner, not on examination of a student record at a later time, as is done for group practical assessment.

Although assessment could be confined to just observed behaviors (e.g., how well a student is able to use a microscope), many tests are developed that also require some interaction between the adult and the child. This may take the form of guidance or prompting, questioning of the student to clarify his thinking or the reason for some action, or having the student respond orally to the task.

For the individual practical task "Food Preferences of Mealworms" (see Figure 5.12), a checklist is provided, which includes specific behaviors to look for as the child goes about finding an answer to the question (Harlen, Black & Johnson, 1981). Low inference, observable behaviors are preferred; for example, "Puts approximately equal quantities of food." There is also a place at which the student can be provided a hint after a few minutes wait if he has trouble proceeding. For example:

> "One way you can try is to make a mark somewhere in the middle of the paper. Take some of the food and place it at the same distance from the mark and then put some mealworms on the mark."

If a hint is given, the student is not given credit for the item, "Deliberately provides mealworms with choice." Nor is the student given credit for the item, "Makes a record of findings without prompting," if prompting is needed.

Laboratory Student Behavior Checklist
Uses hand lens correctly
Hint given
Deliberately provides mealworm with choice; i.e., at least 2 foods at once
Employs an effective strategy such as: (i) uses 6 or more mealworms if all 4 food compared at once (ii) compares foods in all possible pairs with 1 mealworm (iii) tries at least 4 mealworms with one food at a time
Attempts to provide equal quantities of different foods
Puts approximately equal quantities of food
Attempts to release mealworms at equal distance from all foods <u>or</u> arranges mealworms to be randomly distributed around food
Arranges to release mealworms from points equidistant from foods, <u>or</u> places mealworms randomly around foods
Arranges for all mealworms to have same time to choose (i.e., puts them all down together or uses a clock)
Uses clock to time definite events
Allows about 4-7 minutes for mealworms to make choice (not necessarily timed)
Examines behavior carefully (to see if food is being eaten)
Counts mealworms near each pile after a certain time (or notes which food the mealworm is on for strategy (ii) above)
Makes notes at (a) (however brief)
Records details such as time of choice and numbers near each food
Can read stop clock correctly (to nearest second)
Makes a record of finding at (b) without prompting
Results at (a) and (b) consistent with evidence (even if only rough)
Results based on and consistent with quantitative evidence

Figure 5.12. Laboratory Student Behavior Checklist.

At times a child's performance during an activity does not give us a clear indication of his understanding of a procedure or concept. A question asking for justification at this point might shed some light on the reasoning behind an action. For example, a first grader was asked why she kept breaking off smaller and smaller pieces of plasticene in an attempt to get the material to float. Her answer indicated that she was only considering the variable of size. (This is a typical response for this age.) A sixth grader, when questioned about why he added more salt water to a salt water tank, indicated in his answer that he did not know that only the water evaporated, not the whole solution. In both instances a question posed to the child and asking for justification of an action helped in understanding the reasoning behind the behavior.

The oral instructions, hints, and prompts provided in the APU activity, "Food Preferences of Mealworms," were all standardized, thus providing the opportunity for comparisons among the behaviors and responses given by different students. This would lend credibility to any evaluation of students or programs based on these comparisons. Justification questions, on the other hand, sometimes resemble clinical interviews conducted by Piaget and may not be standardized after the first few questions. While they can be useful for evaluation purposes, they are more often used to develop categories of responses to be standardized for future test constructions.

If a student is expected to respond orally to a question, the examiner usually records the response in his notes or makes an audio tape of the interaction. This tape can be used later for a check on the presentation of the task, as well as for evaluation of the student responses. As with any interaction, the examiner must be careful not to unintentionally lead the student to the desired behavior or response. There is a greater tendency to make this mistake if the examiner does not use scripted dialogue.

Materials and equipment

Materials and equipment for an individual practical should be chosen with the same care as those used for a group practical task. It is better if you use materials which are familiar to the student and not complicated. This is especially important if the child is assessed on his skill of using simple laboratory equipment. It is also best to choose materials that do not require a lot of "putting together" by the student. The assembly can only distract from the task.

Although fewer materials might be needed for an individual practical format than for a group practical, it is still necessary to have extras ready in case of an accident or a faulty part. If tests are run consecutively, it is necessary to have duplicate materials for those which can not be reused without some treatment, such as washing. A complete materials and equipment list should be prepared for each activity such as with the mealworms task in Figure 5.13 (Harlen, Black & Johnson, 1981).

Materials List for Student Laboratory Activity

In the "mealworms" investigation the following materials were prepared for the pupil:

About 20 mealworms in a container with no food or sawdust
An additional empty container
A large drawing of a mealworm
A hand lens
4 tubes, filled with bran, sawdust, sugar , and mashed banana
A 30 cm ruler
A spoon (for the food)
A stop-clock
A deep-sided tray lined with plain white paper (approx. 40 x 25 cm)

Figure 5.13. Materials List for Student Laboratory Activity.

The materials that a student is allowed to use for an activity should be in full view of the student as the activity is presented. It should be explained to the student that these are the materials and equipment which he will be allowed to use in the investigation. It is best to limit the student to these items. The materials should be mentioned in the initial instructions for the task, so that they are recognized and, if appropriate, their functions identified.

Organization

Although the organization of an individual practical activity requires careful planning before it is presented to a student, it is not as complicated and elaborate as the planning for a group practical. A place away from distractions and interruptions is needed. An uncluttered workspace is also necessary.

While the time necessary to complete an activity may vary between 20 and 30 minutes, ample time should be allowed for introducing the child to the equipment and for explaining the task. This introduction could add 10 minutes or more to the total time. Although occasionally a child may need encouragement to work more quickly, most children should have enough time to finish at their own pace. Therefore, it is best to leave time between sessions or to have each student notify the next student upon returning to the classroom.

Student Records

A student record sheet is frequently used as part of an individual practical. The record sheet repeats the question or problem that is presented orally to the student and includes a place for jotting down the student's answer or conclusion and any other relevant data. The student record sheet for the APU mealworms task is shown in Figure 5.14 (Harlen, Black & Johnson, 1981). The task is a very open one, and the record sheet provides little structure that would help a student organize the gathering of data. Other sheets might include a chart with the sections labelled. The one for the NAEP water temperature task (NAEP, 1975) is primarily intended for answers and includes a space for computation (see Figure 5.15).

Example of Student Record Sheet

> Find out if the mealworms prefer some of these foods to others. If they do, which one do they prefer?

a) Put down here any notes and results as you go along

b) Write down here what you found about the foods the mealworms prefer

____

____

Figure 5.14. Example of Student Record Sheet.

Example of Student Record Sheet

A. Temperature of the hot water is _____ °C.

B. Temperature of the cold water is _____ °C.

C. I THINK the temperature of the mixture will be _____ °C.

D. Temperature of the mixture is _____ °C.

Use the space below to do any work:

Figure 5.15. Example of Student Record Sheet.

Adaptations

There are also adaptations to this format to suit special conditions and needs. The amount of structure or help provided to the student can be varied, although adding much structure to the task would defeat the purpose of using an individual practical format. For example, if a student were given directions on how to carry out the investigation, then he could not be assessed on the ability to plan or translate a problem into an investigable question. Older elementary school students might present a plan for approval, before being allowed to operationalize it. There has also been some interest in observing *two* students working together on one task. This approach would reveal not only attitudinal information but also reasoning as the two students interact. A complex task could have the number of variables limited for the investigation (e.g., using *one* ball and bouncing it on different surfaces), or a problem with multiple variables could be presented to older, more advanced students. An individual practical could also take place over a longer period of time, and thus include living materials and the time factors needed for growth or change (for example, an investigation of whether seeds need light to germinate.)

Other Opportunities for Assessment

In the previous two sections under "Practical Assessment," assessment situations were described in which a variety of formats were used in combination during one testing session. For example, some required the student to read and respond in writing, while others depended more on the observation of students engaged in practical work. Some involved substantial interaction of the examiner with the student, while others confined the discourse to the management of the test situation. In this remaining section we will discuss some of the specifics, which have not already been explored, of using dialogue, observation, or various products of student practical work for assessment purposes.

Using Interviews for the Purpose of Assessment

The clinical interview approach of Piaget has influenced many people concerned with using dialogue with students to reveal the development of the concepts and mental skills of science. Although the technique may resemble somewhat the Socratic approach used by teachers to guide students toward the understanding of a concept, the purpose differs. During a clinical interview, the interviewer's task is to encourage the child to talk about a topic and then, through adroit questioning, clarify what the student is saying so that the interviewer understands the mental processes and concepts that the child is using. The purpose of the interaction has changed. Instead of leading, the teacher follows so that information may be gathered about where children are in their development or progress.

There is no simple recipe for conducting successful interviews of this sort, for much of the process is interactive. Usually the interviewer starts with a few standard questions prepared prior to the interview. Judgments are then made throughout the discourse to adjust the questions and technique to the student and the situational needs. For example, after a child is invited to roll a car down two ramps which vary in slope, she is asked to predict where the car will go if the slope is increased a specified amount and then to give the reason for her prediction. If the child has trouble with this task, the interviewer can go back to the beginning of the task and ask the child to give observations about the car rolling down the two different slopes. A more advanced student, who had no trouble with the original task, might be asked to state the exact spot to which the car will go and give the quantitative relationship or generalization. Much of the value of the interview method lies in this opportunity to re-phrase, encourage, backup, cue, and generally tailor the test situation to the individual student.

The kinds of questions that are asked in an interview are important and deserving of prior thought. Questions are posed that reveal a student's grasp of a particular concept or the misconceptions that are held about it. Therefore, as the reasoning behind an answer is usually being sought, the questions are framed so that more than a "yes-no" or one-word answer is required. While much value has been attached to those requiring the higher cognitive skills in an instructional setting (using or extending), a mixture of questions is more appropriate for the interview. Questions can be developed also to ascertain how well students use some of the process skills or solve problems. A productive interview depends not only on tailoring the situation to the particular child but also on how skillful the interviewer is in framing questions which elicit a response in the anticipated area.

Some questions might serve to illustrate. "What might be the reason why...?" will tend to elicit answers based on inference. "What do you think might happen if...?" has a good chance of resulting in a prediction. "What is the reason...?" and "What will happen if...?" both will probably result in a wild guess or a response of "I don't know," instead of the desired prediction. However, asking a child for the reason for an observed action or phenomena can be used to probe his understanding of the concept involved. "What do you notice...?" asks for an observation, or a pattern based on an observation, and probably does not need the phrase "do you think" included. "How can we find out...?" invited the child to plan an investigation, whereas, "How would you do it differently if you did it again?" encourages the child to evaluate the strategy he used in an investigation.

Cueing a child to the correct answer, inference, etc. may be a reaction that is hard for the experienced teacher to control. This is no easy task, for along with the habit of using the Socratic approach is the proficiency of many students in "psyching-out" the teacher. A little nod of the head, the tendency to question only the "wrong" answers, and other patterns are soon picked up by the experienced student-turned-teacher-watcher. Cueing, however, is "verboten," as it defeats the major purpose of the discussion.

Keeping accurate records of an interview is of primary importance but difficult and sometimes distracting to both the interviewer and the student if done extensively during the interview. Consider using a tape recorder placed close enough to the child to pick up the voice. Though a hidden microphone could be used, it is probably better to have it unobtrusively visible. The interviewer should explain to the child that it helps her to remember what they have done. You will probably find that they soon forget about its presence. If you think that a particular child continues to be distracted by it and that it therefore interferes with the dialogue, you should make a note of that reaction in your records. The tape will also give you a chance to check on yourself for consistency of questioning over several interviews.

As with the accomplished fisherman, there are certain skills worth developing to increase the chances of catching a particular fish - if it is there. When preparing questions, for example, time needs to be allowed to develop and improve them. Practice sessions can be helpful, too. When and where to use a specific question or technique needs to be left to the judgment of the "fishermen," who will usually increase their skills through experience and practice. The following checklist (Osbourne & Freyberg, 1985) may be helpful for those planning interviews (see Figure 5.16).

A postscript should be added, perhaps, at this point. What has been discussed thus far has focused on interviewing individual children for the purpose of assessment or evaluation. It could be considered an individual competency measure and is usually done with a sampling of students for the purpose of evaluating programs. It can also be used for an individual student or individually for the entire group. However, it would be very time-consuming and not well suited for this purpose or for extensive gathering of data from many individuals. Some programs, such as SAPA, use a similar format in their "group competency measure." However, one can easily be fooled into thinking that the group has attained the competency, unless we are carefully selective about which students we choose to interact with. The technique can be appropriate and effective when trying to determine whether the group has the necessary understanding and skills to proceed with activities building on previously taught skills and concepts.

Observing Students Performing Practical Tasks

The observation of student performance has already been mentioned in the section on individual practicals. It was used in combination with other formats. The technique can also be used alone to assess certain student behaviors during a test situation or normal classroom activities. Behaviors are usually noted on a checklist composed of low inference, observable items.

The technique of observing is most frequently used when skills requiring the manipulation of materials are being assessed. (It would not be appropriate for assessing high-inference behaviors that are not readily observable, such as many of the skills used in reasoning.) How well a student uses a microscope has often been evaluated through teacher observation on the secondary level. Students can also be observed reading a thermometer, using a balance scale, a protractor, a ruler, or most kinds of equipment. The "fairness" of a test may also be observed

Checklist for Interviewers

	Do's	Don'ts
1.	Try to establish clearly how and what the pupil thinks. Emphasize it is the pupil's ideas that are important and are being explored.	Do not give any indication to the pupil of your meaning(s) for the word or appear to judge the pupil's response in terms of your meaning(s).
2.	Provide a balance between open and closed questions and between simple and penetrating questions. In so doing, maintain and develop pupil confidence.	Do not ask leading questions. Do not ask the type of question that is easy for the pupil to simply agree with whatever you say.
3.	Listen carefully to the pupil's responses and follow up points which are not clear.	Do not rush on (e.g., to the next card) before thinking about the pupil's last response.
4.	Where necessary to gain interviewer thinking time, or for the clarity of the audio-record, repeat the pupil response.	Do not respond with a modified version of the pupil response; repeat exactly what was said.
5	Give the pupil plenty of time to formulate a reply.	Do not rush, but on the other hand, do not exacerbate embarrassing silences.
6	Where pupils express doubt and hesitation encourage them to share their thinking.	Do not allow pupils to think that this is a test situation and there is a right answer required.
7.	Be sensitive to possible misinterpretations of, or misunderstanding about the initial question. Where appropriate explore this, and then clarify.	Do not make any assumptions about the way the pupil is thinking.
8.	Be sensitive to the unanticipated response, and explore it carefully and with sensitivity.	Do not ignore responses you do not understand. Rather follow them up until you do understand.
9.	Be sensitive to self-contradictory statements by the pupil.	Try not to forget earlier responses in the same interview.
10.	Be supportive of a pupil querying the question you have asked, and in this and other ways, develop an informal atmosphere.	Do not let the interview become an interrogation rather than a friendly chat.
11.	Read the question out loud to pupils.	Do not rely on pupils' reading ability.
12.	Where all efforts to develop pupil confidence fail, abort the interview.	Do not proceed with an interview where the pupil becomes irrevocably withdrawn.
13.	Verbally identify for the audio-record, the pupil's name, age, and each card as it is introduced into the discussion.	Do not return to earlier cards without verbal identification for the audio-record.
14	Be sensitive to the possibility that pupils will give an answer simply to fill a silence.	Do not accept an answer without exploring the reasoning behind it.
15.	Appreciate that a card omitted will result in missing data.	Make no assumption about the way a pupil would respond to a particular card.

Figure 5.16. Checklist for Interviewers

as students go about conducting an investigation. For example, whether everything but the independent variables are kept the same, or controlled, can be observed. Behaviors not requiring interaction for the purpose of clarification or justification can be assessed through observations.

Which skills, and how many, are observed in any single session would depend on the particular activity, the time available, and the number of students being observed. If children are working independently with repeated readings of the temperature of liquids, then a whole class might conceivably be assessed for their ability to use thermometers properly. On the other hand, the type of skill and the amount of time available might determine that a fewer number of students would be assessed.

Developing a checklist for a particular skill is usually not a difficult task for an experienced teacher. It can often follow rather closely the instruction given to children, as well as the mishaps that often ensue. A checklist intended for third graders on the use of the thermometer might look as follows:

- holds thermometer so that thumb is not on bulb
- places the bulb part of the thermometer on or in the material to be tested
- reads thermometer while it is on or in the material
- waits a suitable duration until the temperature reading stabilizes
- uses the same thermometer on subsequent readings or uses another thermometer that has been compared with the original one for accuracy
- does not attempt to "shake it down" if it is not a fever thermometer

If the observable steps in conducting a fair test to answer a particular question are listed, then a checklist can be used for assessing behaviors, such as controlling all but the independent variable, measuring or weighing with care, recording, and many other behaviors similar to those described in the section about individual practicals.

Although assessment by observation might seem to be too time consuming, many important laboratory skills cannot be evaluated otherwise. For example, only one aspect of using a thermometer can usually be tested by paper and pencil means; i.e., the ability to read a number line. As with any other assessment method, assessment by observation becomes easier with practice.

Using Student Writing, Projects, and Other Products for Assessment

There are a number of products that children create as a result of their interaction with scientific phenomena and materials. These products include projects typically done at the end of a unit of study and much of the writing, charts, graphs, and illustrations done during the learning experience. Many of these student creations indicate what the student knows and can do. The procedures used with these products to assess student understanding and skills might include some adult-child interaction similar to an oral examination or the assessment might rely totally on the adult's evaluation of the product.

After a student completes a project, it is often presented to the class or to other groups. Although the project might "speak for itself" and indicate the level of skill and understanding that the creator has attained, an oral presentation or a question and answer period can contribute greatly to the assessment. The presentation could begin with a brief, unstructured explanation of the project by the student and then have questions from an examiner. These questions would deal specifically with the project and focus on the skills and concepts of the project's topic area.

Questions for this purpose are developed with the same considerations described in the section "Using Interviews for the Purpose of Assessment." A few standardized questions which are generic enough to apply to all projects but also easily made specific to the particular topic will provide focus to the discussion and will allow for comparisons to identified criteria or norms. For example, if children have raised a variety of organisms as a project, questions could be asked as follows: how is a particular structure related to the function it serves; how could an observed behavior be a helpful adaptation; and what might be the effect (if any) if the population of the particular organism suddenly increased?

An interesting "first cousin" of the final project assessment procedure is the method used by the Science Olympiad. That method requires students to apply skills and concepts to a problem-solving situation. For example, the activity entitled "Aerodynamics" in Figure 5.17 involves constructing a paper airplane and flying it at a target (Science Olympiad, 1986). Two students work as a team. It is assumed that the most successful students will be those who have the best understanding of the phenomena and can apply the appropriate concepts and skills to the situation. The scoring guide indicates that the team whose planes come closest to the target wins.

The procedure for assessment, used in the previous examples, involved either adult-student interaction or observation. Examination of just the student's writing, project, graph, illustration, etc. without added dialogue or observation is a method that is also frequently used in assessment. This procedure would be familiar to elementary school teachers who have been involved in methods currently used to mark children's writing in the language arts area of the curriculum. Sometimes, this involves a rather general appraisal. Other times, a comprehensive assessment is required. Still other situations call for gathering data on only a limited number of skills or concepts evidenced in the student's creation. What kind of information is collected from various student records will depend on the objectives for the assessment. This often depends in part on how much time is available for the assessment and, thus, on how comprehensive the assessment is.

In Chapter 4 a model of problem-solving was presented which describes the basic stages of planning and conducting an investigation or fair test (see Figure 4.2). All or any part of this model could be used for assessment. For example, if the teacher wants only to determine whether the student has a general idea of

Example of Guidelines for Scoring a Student Competition

Aerodynamics

Each two member team will build one paper airplane to be flown a distance of at least five meters, landing on a predetermined target. Airplanes must be of a folded aerodynamic design. Crumpled wads of paper do not qualify.

Number of participants: 2 Approximate time: 60 minutes

The Competition:

1. Two sheets of plain white paper (ditto paper) will be supplied for each team along with approximately five centimeters of masking tape and a pair of scissors.
2. Planes flown in competition must be made on site, during the allotted time, using only the materials provided.
3. Planes will be hand-launched from behind a line on the floor at a specified target, on the floor, more than five but less than ten meters distant.

Scoring:

1. After the flight, the distance will be measured from the center of the target to the nose of the airplane where it comes to rest. The distance from the target will become the school's score.
2. Each team member will fly the plane once. The team score will be determined by adding the two scores.
3. The lowest score, signifying the closest to the target, will be the winner. In the case of a tie, there will be a "fly-off."

Figure 5.17. Example of Guidelines for Scoring a Student Competition.

the investigation procedure and outcome, then a quick reading of the report can indicate areas of strength and weakness. The simple report in Figure 5.18 might indicate to the reader that this third grader (Maki, 1985) had a good grasp of what happened during the "fair test" and of what the outcome was. This type of "first impression" assessment is usually referred to as "holistic evaluation."

On the other hand, a comprehensive assessment of the report of an investigation might include all of the categories and sub-categories of the model appropriate to the particular activity. (Some behaviors under the performance category might be inferred, but most are better assessed through observation.) Obviously, such a process would be very time-consuming when marking the papers and when

Example of a Student Report

Science
Mrs. O'Connor
January 28, 1988

Yesterday at Science I got a container just about the size of a regular ice cube. Everybody in the class started at 12:29. Other smaller container icecubes melted faster then mine and bigger ones took longer to melt.Now I know that it take longer to melt if it is bigger.

Figure 5.18. Example of a Student Report.

constructing a scoring guide for each of the numerous sub-categories. Also, consider the poor elementary school-aged children who would have to "push the pencil" to provide you with all of the material from which you would glean the desired information. While keeping records in science is a skill that every child should master, it is not the "end all and be all" of the activity.

Far more common is the use of only part of the rich array of potential information in a student report. This is called primary trait evaluation. One or a few sub-categories of the model are identified for assessment. For example, it might be of interest to a teacher to determine if the students could gather and organize the data in an appropriate manner. In addition, it might be appropriate to the teacher's objectives to see if the students could then gain information from the graph that they had constructed. A scoring guide, which is written expressly for the particular activity on which the student reports, would include only criteria for the sub-categories under assessment. For example, in a situation in which students make and use a graph of comparative plant growth under conditions of darkness and light, the teacher might look for:

- a line graph (not two separate bar graphs)
- an appropriate title reflecting the data (e.g., "Comparison of Growth of Grass in the Dark and with Light")
- both axes labelled appropriately (height and date)

- appropriate scale
- accurately made from data (no data recorded on graph incorrectly)
- correct summary based on data in graph
- conclusion consistent with the data presented in graph (not necessarily "right")

It is not uncommon for elementary school children to gather data and then come to conclusions that don't relate to the data. For example, a boy measured the *distance* a car went on various slopes and then concluded that the steepest slope caused the car to go the *fastest*. Another student accurately recorded the difference in insulating quality between two materials but then was unable to conclude from the graph which material was the better insulator.

Student records of science experiences, such as charts, graphs, diagrams, and writings, can be used in assessing progress. How much information and what kind of information can be gathered from those records depends, however, on the richness of the experience and the specific assignment the child is given. An activity which requires the student to use a variety of skills and concepts is richer than one in which a student might be asked only to observe and identify differences. A student's recording of the experience can vary also. Sometimes the "specific assignment" is teacher expectations built over time, or student experiences with what and how the teacher has marked previous papers. For example, the astute student would be foolish to record only what was found out from an activity, when the teacher regularly marked papers on how well the student described his experimental procedure. On the other hand, "lab sheets" are often so inclusive and directive that the blanks left for the children to fill in yield information only on how well students can follow directions. The assessment potential using these student products and records, therefore, is restricted by the nature of the activity and the student's understanding of the recording assignment.

CHAPTER 6

USING THE INFORMATION GATHERED

This chapter is devoted to the techniques for using the information gathered for the purpose of evaluating students and programs. It is organized into several sections: scoring the responses, interpreting and presenting the findings, grading and evaluating students, and program evaluation.

Scoring Written Test Items

The scoring of student responses will depend largely on the nature and format of the questions used. For most written test items, it is a simple process of adding the number of responses which agree with the response keyed "correct." The scoring (marking) scheme is such that the student gets one point if he checked the "keyed" choice as his answer and no points for any other answer. Students also receive no credit if they have indicate two or more answers as being correct.

There are test items in which more than one choice is expected for full credit to the item. The item illustrated in Figure 6.1 (Harlen, Palacio & Russell, 1984) asks the students to respond to *each* of the choices - indicating by putting a tick in the appropriate box whether "it is something they would find out and a cross if it is not." A maximum of four points is available for this item, one point for each of the four "correct" responses.

A common complaint of teachers and researchers is that one seldom knows *why* a student chooses a particular answer. There are those who would argue that the reason is more important than the choice. In the item cited in Figure 6.2 (Harlen, Palacio & Russell, 1984), one point is earned for choosing the correct choice (30 years) and two points for the reasoning given for the answer. There is partial credit of one point for a "correct but restricted statement" explaining the answer. Teachers could categorize the answers which students provide for the reasons and develop key descriptors for grading subsequent groups of students. The number of points allotted for the reasons could be more than two if there is a greater diversity of answers which fit into some pattern for distributing points.

The scoring system described above will also work well with other "objectively scored" forms of test items; e.g., true/false and matching. For completion, short answer, and essay questions the task is a little more complicated but is greatly

<table>
<tr><td>

Example of Multiple True/False Items

Some girls had a collection of toy cars of different colors and sizes.

This is the investigation they were doing with them.

1. Beth and Eve stood at one end of the playground and Joan at the other end
2. Beth and Eve put different colored large cars on the ground one at a time.
3. Joan tried to tell the color of each car as it was put down.
4. Then they did the same with the different colored small cars.

What would they find out from this investigation?

Read the sentences below. For each one put a plus (+) if you think it is something they would find out and a zero (0) if it is not.

☐ Which colors could be seen from a distance.
☐ Whether the distance made any difference as to which colors could be seen.
☐ Whether Joan was better at seeing the colors than Beth and Eve.
☐ Whother the size of the car made any difference as to which color could be seen from a distance.

</td><td>

Marking (maximum score = 4)

Response	Marks
☐	1
☐	1
☐	1
☐	1

</td></tr>
</table>

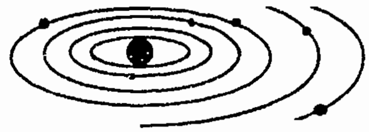
Figure 6.1. Example of Multiple True/False Items

simplified by the preparation of a detailed "model answer." These model answers will include sets of expected answers, as well as information for partial credit. If the item is well constructed, the preparation of this is quite easy. The item in Figure 6.3 (Harlen, Palacio & Russell, 1984) has a maximum score of three points, one point each for three distinct responses that are obvious elements of the test question .

In science, many activities and test items include a graphical, tabular, or pictorial presentation. The ability to construct and interpret information from these modes is stressed in elementary science programs. The item in Figure 6.4 (Harlen, Palacio & Russell, 1984) requires students to make several observations from the graphical information provided. A total of four points is available for the item, one point for each of the three questions and one point for the inclusion of the correct units of measurement with each response

Example of a Scoring Guide for a Test Item

Planets move around the Sun



Look at the following table.

Planet	Distance from the Sun	Time for one trip around the Sun
Mercury	68 million kilometers	88 days
Venus	108 million kilometers	225 days
Earth	150 million kilometers	1 year
Jupiter	780 million kilometers	12 years
Uranus	2,870 million kilometers	84 years
Neptune	4,500 million kilometers	165 years

a) There is another planet not in this table. It is about 1,430 million kilometers from the Sun.

About how long do you think it will take this planet to make one trip around the Sun?
Mark the box next to the one you choose.

☐ 10 years

☐ 100 years

☐ 100 days

☐ 30 years

☐ 300 days

b) Why do you think it will take this time?

Because________________________________

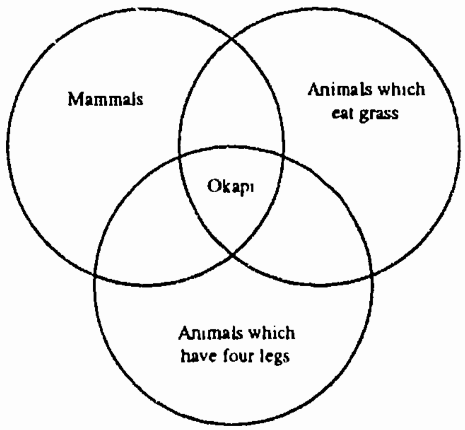
Marking
(maximum = 3)

Response	Mark
a) 30 years marked	1
Anything else marked or multiple marks	0
b) Statement that there is a connection between distance and time so that if the distance is between Jupiter and Uranus so must be the time	2
Correct but restricted statement which omits overall pattern, e.g., "It is between Jupiter and Uranus." Statement implying relationship between time and distance only.	1
No informative answer, e.g., "It fits" or irrelevant or incomprehensible	0

Figure 6.2. Example of a Scoring Guide for a Test Item

The item illustrated in Figure 6.5 is an example of the "partial credit" allotted to each element of an answer (Harlen, Palacio & Russell, 1984). In this case, three important bits of information would comprise the "totally correct" answer. Notice also the need to accept synonyms for some of the words; e.g., "goes into" for "dissolves."

Example of a Scoring Guide for an Essay Test Item



Write down all that you can tell from this diagram about what sort of animal an okapi is:____________

____

____

____

Marking
(maximum score = 3)

Response	Mark
Okapi	
is a mammal	1
eats grass	1
has four legs	1

Figure 6.3. Example of a Scoring Guide for an Essay Test Item

In the graphing item in Figure 6.6, it is clear that producing an adequate graphical presentation is a complex task (Harlen, Palacio & Russell, 1984). A total of 15 points is available for this item - one point for each of 15 discrete elements. Five of these points are allocated to the labelling and scaling of one axis and another five points for the other axis. Some teachers would expect that the vertical axis present the data for the dependent variable (in this case, the height of the bean plant). However, in this scoring scheme the data could be plotted on either axis. Lastly, five points could be earned for the correct entry of the five "data points" on the graph.

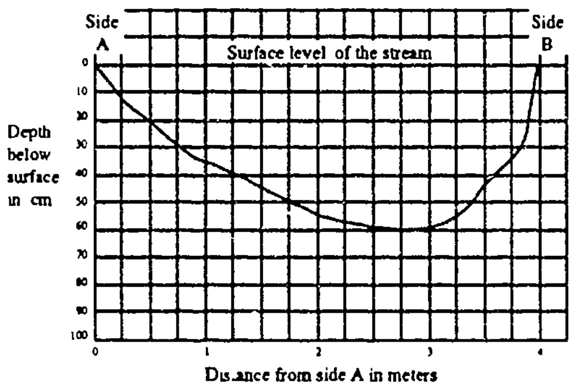
Scoring Practical Tests

The scoring of group-administered tests can be accomplished in a number of ways. Most tests use a scoring system with point values associated with particular observations, measurements, or statements. One of the most difficult process tasks to assess is the Planning and Design of an Investigation. Figure 6.7 illustrates a "general scheme" that might be useful for evaluating student efforts at planning (Lunetta, Hofstein & Giddings, 1981). There are other samples later in this section, with detailed analyses of necessary elements for assessing skills with particular investigations.

Example of a Scoring Guide for a Graphing Item

Some children measured a stream. They called one side "side A" and the other "side B." They found how wide the stream was. They also found out how deep the water was below the surface.

They made a graph with their results.



Use the graph to answer these questions.

a) How wide is the stream?

b) How deep is the stream at the deepest point?

c) How far from side A would you need to go to get a depth of 35 cm?

Marking (maximum score = 4)

Response	Marks
a) 4 m.	1
b) 60 cm.	1
c) 1 m. or 3.75 m. Accept 90-110 cm.	1
Correct units given for each response.	1

Figure 6.4. Example of a Scoring Guide for a Graphing Item

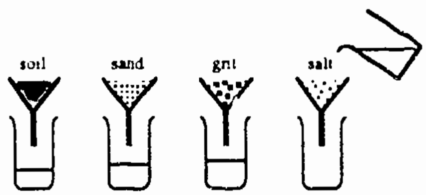
One of the most common tasks in group practical testing is the observation of similarities and differences. The stimuli for the observation task can be actual objects, live specimens, models, drawings, or pictures. The example in Figure 6.8 refers to two insects, a spider and a crane-fly (Harlen, Black & Johnson, 1981). The scoring scheme allocates three points (maximum) for descriptions of ways in which they are the "same" and three points (maximum) for listing three ways in which they are "different." The list of acceptable responses may have to be increased as bright, creative students respond to these questions. A key element to the marking scheme is that statements must describe "a feature which could be *observed* from the specimens, real or simulated, not based on information which might well be true but was recalled rather than observed..." This is a fair procedure because the phrase "from the drawings" was present in the question and underlined. Further, note in the "comments" section that a phrase like "both have..." is sufficient for a statement of similarities but that "they had different..." was not an acceptable description of a difference. Students were required to state "how one thing was different from the other"; e.g., "Crane-fly has wings, spider does not have wings." ("Crane-fly has wings" was not enough.)

Example of a Scoring Guide for an Interpreting Test Item

David and John put equal amounts of dry sand, soil, grit, and salt in four funnels.

They wanted to find out how much water each one would soak up. So they poured 100 ml of water into each one.

This worked all right until they came to the salt. When they poured the water in, almost all of the salt disappeared.



Why do you think the salt disappeared but the other solids did not?

I think this might be because______________
____
____
____

Marking
(maximum score = 3)

Response	Marks
Salt dissolves in water (allow "goes into" or similar).	1
Salt carried away in the water.	1
Soil, sand, and grit do not dissolve in water.	1

Figure 6.5. Example of a Scoring Guide for an Interpreting Test item

Another common activity of the early grade levels is the identification of objects by using the senses. In Figure 6.9 the sense of touch was used as students had to touch (without looking) five materials and determine which one was rubber (Harlen, Black & Johnson, 1981). They were told which five materials were in the box. One point was earned for the correct identification and up to two points more for a description which included properties which could distinguish rubber from the other materials, such as soft, can be compressed, etc.

The following illustration (Figure 6.10) is from the task in which the students were to measure and record the temperatures of samples of hot and cold water and then predict and measure the temperature of the water when mixed. Note that no points were awarded for the measurement of the initial temperatures, as they could vary from school to school and vary during the test administration period. This task was used in the U.S. testing as part of the Second IEA Science Study (SISS, 1986

Example of a Scoring Guide for Graphing Skills

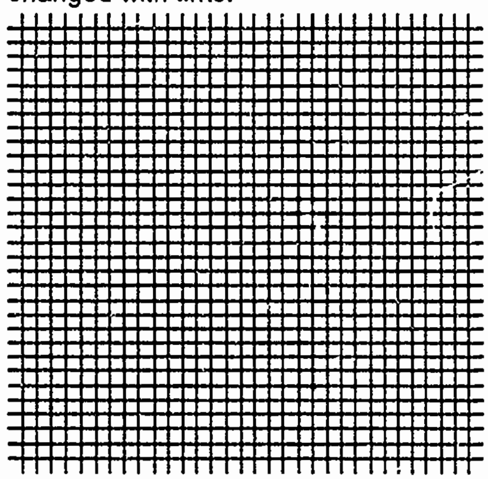
Richard measured his bean plant every week so that he could see how fast it was growing.

He started (0 weeks) when it was just 5 cm high.

These were the heights for the first 4 weeks:

0 weeks	-	5 cm
1 week -	15 cm	
2 weeks	-	30 cm
3 weeks	-	40 cm
4 weeks	-	45 cm

Draw a graph to show how the height changed with time.



Marking
(maximum score = 15)

Response	Mark
One axis	
Labelled 'Height', How Tall', 'Bean Height', etc.	1
Labelled with name of units (cm).	1
Equal interval scale.	1
Suitable scale.	1
Scale labelled (numbers attached to divisions).	1
Other axis	
Labelled 'Time', 'How Long Growing', etc. (n.b. not 'How Fast').	1
Labelled with name of units (weeks).	1
Equal interval scale.	1
Suitable scale.	1
Scale labelled (numbers attached to divisions).	1
Drawing	
0 weeks - 5 cm	1
1 week - 15 cm	1
2 weeks - 30 cm	1
3 weeks - 40 cm	1
4 weeks - 45 cm	1

Figure 6.6. Example of a Scoring Guide for Graphing Skills

The idea of a "fair test" (see Chapter 4) has been used widely in helping elementary school students develop their skills in designing investigations. Some elements of the "fair test" can be handled by youngsters in third grade. In the item in Figure 6.11 (Harlen, Palacio & Russell, 1984) the students are asked to list three things that should be kept the same for both walls in this test to see which is stronger. A maximum of three points can be earned through this item by listing any three statements from a long list of acceptable responses. A shorter list of "unacceptable" responses was also provided. Lists like these can be developed through repeated interactions of teacher and students with such activities.

At a higher grade level students are asked to plan a test for determining which wood to use as a chopping board (Figure 6.12). The students are reminded to include "which things you would use," "what you would do," and "how you

Assessment Criteria for Planning and Design

Criteria	Score
Able to present a perceptive plan for investigation. Plan is clear, concise, and complete. Able to discuss plan for experiment critically.	9-10
Good, well-presented plan, but needs some modification. Understands overall approach to problem.	7-8
Plan is O.K., but some help is needed. Not a very critical approach to problem	5-6
Poor, ineffective plan needing considerable modification. Does not consider important constraints and variables.	3-4
Little idea of how to tackle the problem. Much help needed.	1-2

Figure 6.7. Assessment Criteria for Planning and Design

would find out the result." This task has been analyzed by the British APU researchers (Harlen, Palacio & Russell, 1984). They have determined ten categories on which student responses can be evaluated. A total of seven points can be earned for a fully acceptable answer. The points are earned in the following categories:

General Approach	1 point
Number of Blocks Used	1 point
Variables - Tests	2 points
Relevance of Tests to Problem	1 point
Measurement/Observation of Result of Each Test Given to Blocks	1 point
Interpretation and Recording of Results	1 point

Some of the categories are not used for allocating points but for describing the nature of the student's response (Figure 6.13). In Figure 6.14 a student answer and its corresponding evaluation is listed (Harlen, Palacio & Russell, 1984). Two important elements of these evaluations are whether the statement was explicit and whether something was clearly identified or whether it was assumed.

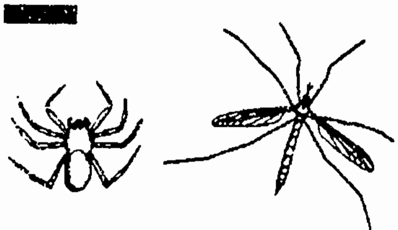
The scoring of student answers to process test items should be done by someone other than the classroom teacher of the students being assessed. As teachers, we are prone to "read into" the answers more than the words really indicate. A school system planning to use these tests over several years would be wise to train a team of people to do the scoring. It is necessary to allow sufficient time to train the scorers thoroughly (minimum of six hours). Possible procedures to train the team to score the practical tests will be discussed briefly.

Training should begin with a familiarization with the specific tasks and scoring being used. Then a set of completed test booklets should be scored by each member of the team. Differences in scores and the evidence used will be

Scoring Guide for Classification Activity

Question Page

[redacted]



Look at the drawings of a spider and a crane fly.

Find three ways in which you can tell from the drawings that they are the same as each other and three ways in which they are different from each other.

Same	Different
1 ______	1 ______
2 ______	2 ______
3 ______	3 ______

Comment

To qualify as an acceptable response a statement of similarity or difference had to describe a feature which could be observed from the drawings, not based on information which might well be true but was recalled rather than observed, such as the spider spins a web but the crane-fly does not. In some cases the line was hard to draw since an answer that the crane-fly is an insect and the spider is not, could be made on detailed observation that the crane-fly had all the characteristics of an insect whilst the spider did not. However it was decided that only statements about specific features would gain marks. By the same rule credit was not given for the statement that the spider could not fly whilst the crane-fly could, whilst it was given for stating that the crane-fly had wings and the spider had none.

The mark scheme (see below) gave equal credit for each statement of similarity or difference but it was somewhat easier to state a similarity than to state a difference. For the latter it was not enough to state what was different but it was necessary to state how one thing was different from the other. Thus to state 'different legs' was not a sufficient answer for a difference, whilst 'both have legs' was sufficient for a similarity. Note that the mark scheme gives examples of types of answers gaining marks but these were not the only answers allowed.

Mark scheme

One mark for each acceptable point.

Examples:

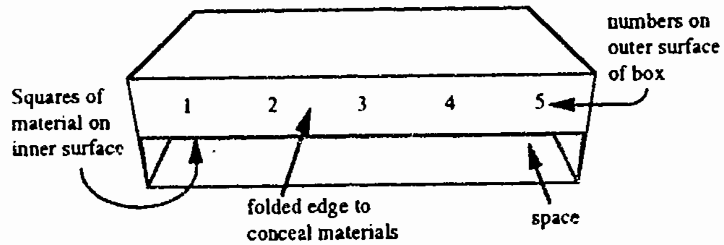
Same (Max. 3 marks)	Both have long thin legs. Both have jointed legs. Both have things sticking out of one part of head. Both have legs sticking out of one part of the body.
Different (Max. 3 Marks)	Spider is hairy, crane-fly is not. Crane-fly has wings, spider doesn't. Spider has fat body, crane-fly's body is long and thin. Crane-fly has point at end of body, spider is round.

Maximum = 6 marks

Figure 6.8. Scoring Guide for Classification Activity

Assessment Guide for a Sensory Activity

Material given to pupils:



Question Page

Inside the box behind each number there is a thin square which may be made of:

glass
metal
wood
leather
rubber

Put your fingers in the box and feel the squares.

(a) Decide which is <u>rubber</u>.
Write down the number in front of the one you think is rubber:

☐

(b) How did you decide it was this one?

____

____

____

____

Comment

The tester showed the pupils how to put their fingers into the box through the space and touch the squares of material on the inside of the numbered surface. The tester then presented the question in the following words:

For this question you use your finger-tips to feel the surfaces inside the box behind the numbers, like this. Stuck on the inside there are thin squares of five different things: glass, metal, wood, leather, and rubber. By feeling them only (don't try to peek!) decide which one is rubber. When you have decided, put down the number which is in front of the one you think is rubber. Then write down at (b) how you decided it was this one.

Marking scheme

a) 5 — 1 mark
b) One mark for each acceptable property to a maximum of — 2 marks
Examples of acceptable properties:
smooth
soft/squashy
spongy/bouncy
can press into it

To be acceptable the property must be one which could distinguish rubber from the other materials.

Maximum = 3 marks

Figure 6.9. Assessment Guide for a Sensory Activity

Scoring Guide for Laboratory Exercises		
Item Number	Answer	Scoring
1	The water should be about 0°C (verified by the test administrator).	No points awarded.
2	The water should be at 50°C. (Or temperature verified by administrator. Water will cool during testing process.)	No points awarded.
3	The water temperature should be predicted to be at mid-point (of temperatures recorded in #1 and # 2).	1 pt for correct prediction.
4	The water temperature should be at mid-point (of temperatures recorded in #1 and #2).	2 pts if the recorded temperature is within 6°C of the calculated mid-point. 1 pt if the recorded temperature is within 10°C of the calculated mid-point.
5	Plausible explanations: hot water cooled excessively, cold water warmed excessively, some water was spilled, volume of water in one cup may have been greater than in the other cup	1 pt for a plausible explanation.
6	The water should be at 40°C. The temperature of a mixture (of equal amounts) will be at the midpoint between the initial temperature values.	1 pt if recorded temperature is correct, or if explanation is correct, or both.

Figure 6.10. Scoring Guide for Laboratory Exercises

discussed. A "master" set of answers is obtained for the set of test booklets from the specialists who developed the test items. This "master" set of scores is used as a template to gauge the success of the scorers being trained. A level of agreement with the "master" set of at least 90% must be consistently reached before scorers can be considered fully trained.

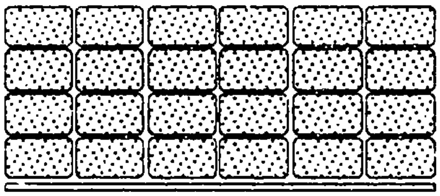
The training of observers for individual practical tests presents unique needs. The training must be specific to each task to be observed. The observers must have a solid understanding of the activity and comparable students being observed. If videotape facilities are available, a novice could score the tape of experienced observers until 90% inter-rater agreement is reached. Similarly, the tape and scoring of a novice can be verified by an experienced administrator. If relatively few students are to be tested, it may be wise to train an observer for each task—creating highly competent specialists for each task.

Interpreting and Using Results

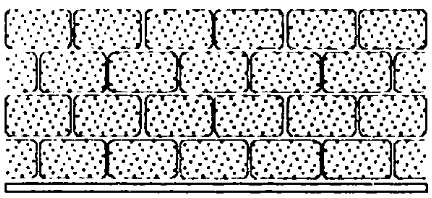
We will discuss several ways of using the results from both written and practical test formats. Obviously, the nature of the reports and the kinds of interpretation possible will be dependent on the nature of the information

Scoring Guide for Planning a Fair Test

Tim built a 'wall' using boxes as bricks, like this:



Flo built her wall from the same kind of boxes, like this:



They wanted to see which wall was the stronger. They decided to do this by rolling balls along the floor to hit the walls. To make this a <u>fair test</u> they must keep some things the same for both walls.

Write down <u>3</u> things that you think should be kept the same:

1.

2.

3.

<u>Marking</u>
(maximum score = 3)

One mark each acceptable different response to a maximum of 3.

Examples of acceptable responses:

Balls
- roll from the same distance away from the walls,
- roll with the same amount of force/same speed,
- test in the same place/room for both walls,
- hit the walls in the same position,
- use the same ball/ weight of ball/size of ball,
- same number of balls used.

Walls
- same height,
- same number of boxes,
- same height of boxes,
- boxes same distance apart,
- same thickness of walls.

Examples of unacceptable answers:
- build walls in same way,
- bricks in same pattern,
- instructions (e.g., do it three times),
- alternative tests (e.g., sit on them).

Figure 6.11. Scoring Guide for Planning a Fair Test

available. With the results from written test items, a very detailed kind of statistical report might be appropriate (see Figure 6.18). This report illustrates what is useful for teachers, evaluators, and curriculum developers. It provides detailed information about the relative success of specific items and identifies prevalent misconceptions and errors. The basic format of the report is a list of percentages of students that choose each of the responses. The "correct" choice is indicated by an asterisk or some other easily recognized symbol. The percentage of students that choose a correct answer is an important bit of

Student Sheet for Planning an Investigation

Suppose you are going to make a chopping board to use for cutting bread or chopping vegetables or meat. You have to decide which is the best kind of wood to use. You have blocks of four different kinds of wood (A, B, C, D) and you can use any of the things in the list below to do some tests on them. (You don't have to use all the things.)

What would you do to:

Test the blocks to find out which kind of wood is best for making a chopping board.

List of Materials:

4 blocks (A, B, C, D)	Nails
Heavy steel ball	Hammer
Butter	Dropper
Knife	Water
Felt-tip pen	Paper towel
Drawing pins	

Make sure you say.

- which things you would use
- what you would do
- how you would find out the result

Figure 6.12 Student Sheet for Planning an Investigation

information. This value has historically been called the "difficulty" index, although it is really an "ease" index. This index is simply the percentage of students tested that give the correct choice. One can describe an item with a range of descriptors from very easy to very difficult. The following categories might be useful for this purpose:

85% - 100%	Very Easy
70% - 84%	Easy
50% - 69%	Moderate
35% - 49%	Quite Difficult
Below 35%	Very Difficult

These criteria will vary somewhat depending on the number of choices used and the approach of the assessment.

Additionally, one can note the percentage of students that choose each of the distractors. As distractors, the function of these choices is to attract some of the students. When the percentages of students choosing some of the distractors

Scoring Guide for Planning an Investigation

Assign marks to those categories where (1) or (2) is indicated to give a total out of 7.

1. General approach

Judge from 2 or more tests applied to all blocks (explicit)	A(1)
Judge from one test applied to all blocks (explicit)	B(1)
Judge from one or more tests applied to all blocks (implicit)	C(1)
Judge from 2 or more tests applied to unspecified number of blocks or "the wood"	D
Judge from one test applied to a block/piece of wood	E
Find out by making board/using it	F
Statement of answer (X is best) - no tests suggested	G
Consult carpenter/expert/other source of information	H
Irrelevant (e.g., how to make a chopping block)	J
No response to any part of the question	N
Different tests to different specified blocks	K

2. Number of blocks used

All blocks used for all tests	A(1)
Some elimination after first or a subsequent test	B(1)
"Blocks" (unspecified) tested	C
One block/the block/the wood tested	D
Chopping board tested	E
No mention of blocks but use implied	F
No mention of blocks and no tests suggested	G
Blocks (specified) tested	H

3. Equipment

Fully specified (identifiable) and from that given	A
Fully specified but not all from that given (e.g., meat to cut up)	B
Not specified but use implied	C
No equipment mentioned nor use implied	D

4. Number of tests (regardless of nature and relevance)

1 test	A
2 tests	B
3 tests	C
4 tests	D
5 or more tests	E
No tests	F

5. Variables - tests

Same amount of treatment specified for each block	A(2)
General statement that all blocks tested in the same way (e.g., "do the same for all blocks") but control not specified	B(1)
No indication that blocks all tested in the same way	C

Figure 6 13A. Scoring Guide for Planning an Investigation

Scoring Guide for Planning an Investigation (Continued)

6. Variables - blocks

Mention of using similar surfaces of blocks for tests	A
Attention to other block variables before testing (cleanness, etc.)	B
No mention of using similar surfaces or other block variables	C

7. Repetition

Mention of repeating all tests using different areas of the same blocks/wood to check result	A
No mention of repetition of tests	B
No test	C

8. Relevance of tests to problem

3 or more tests:	
All relevant and relating to different properties	A(1)
All relevant but some redundancy	B(2)
Relevance of some unclear	C
2 tests:	
Both relevant and related to different properties	D(1)
Both relevant but testing same property	E(2)
Relevance of one or both unclear	F
1 test:	
Relevant to problem	G
Relevance unclear	H
No tests described	J

9. Measurement/observation of result of each test given to blocks

Details given of measurements made to assess results	A(1)
Details given of what to look for to assess result (qualitative comparison)	B(1)
Vague statement, such as "see which is best"	C
No indication of how to find result	D

10. Intepretation and recording of result (to find which is best)

Put blocks in rank order for each test and combine rankings	A(1)
Mention of property used as basis for judgment ("which is strongest")	B(1)
Find "which is best" for each test (when more than one) but no indication of how to combine to give one final result	C
Find "which is best" when only one test	D
Find result "by comparing" (property used as basis not specified) or other vague indication that results follow from tests on several	E
Result obtained from testing one block/it/the wood	F
Mention only that the result would be recorded: no mention of how it is obtained	G
No mention of how result obtained form observations or recorded	H

Figure 6.13B. Scoring Guide for Planning an Investigation

Example of a Student Answer with Codes from Scoring Guide

Make sure you say

- which things you would use;
- what you would do; and
- how you would find out the result.

I would use the hammer and a nail to find out which is the strongest piece of wood. I would hammer a na into the wood with two hits and see what piece of wood would make the shallowest hole then I would use that piece of wood for the chopping board.

Codes assigned from FIGURE 6.16	
1.	B (1)
2.	C
3.	A
4.	A
5.	A (2)
6.	C
7.	B
8.	G
9.	B (1)
10.	B (1)

Figure 6.14. Example of a Student Answer with Codes from Scoring Guide

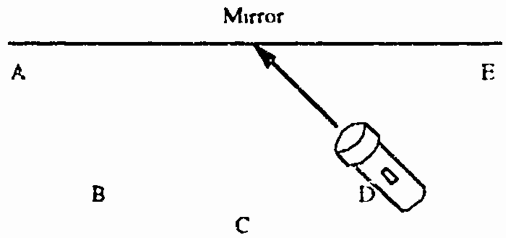
exceeds 10%, these questions may represent misconceptions or error patterns worth examining, or clues within the distractor. One can learn much about an item from this information (Figure 6.15). One can also compare sets of students, such as boys and girls, simply by calculating separate sets of values for each group (Jacobson & Doran, 1985).

Another method of presenting results is through a bar graph. The item and results illustrated in Figure 6.16 (Harlen, Black & Johnson, 1981) are for a short-answer question, with an attached reason for the answer. Over 70% of the students earned two points (of the maximum three points). This is most likely a result of a correct answer (between 14 and 19 days) and a one-point reason, one which does not mention the pattern or connection between the size of the eggs and the time to hatch.

For the item illustrated in Figure 6.17 (Harlen, Black & Johnson, 1981) the students are required to sum the data provided by "number of lengths" and produce a bar chart of the corresponding height. Of the maximum four points for that item, three points are earned by the correct entry value for the length categories: "can't swim," "1 length," and "2 lengths." The fourth point is earned by correctly drawing vertical lines at equal intervals on the base to form the bars. Notice that over 30% of the students earned the maximum four points. Approximately 30% earned no points and about 10% received one, two, and three points.

Another graphing item is illustrated in Figure 6.18 (Harlen, Black & Johnson, 1981). In this item the students must label the axes, determine the appropriate scale interval, and enter the five data points. A maximum of 9 points

Response Distribution for a Test Item by Gender



Your friend shines a flashlight at a mirror, as shown above. Where should you stand so that the light will be reflected onto you?

A. Position A
B. Position B
C. Position C
D. Position D
E. Position E

	A	B	C	D	E	OMIT
Boys	4.5	60.9	22.4	7.7	2.6	2.0
Girls	5.7	34.5	37.8	13.9	4.7	3.4

Figure 6.15. Response Distribution for a Test Item by Gender

is allotted to this item, with partial credit as shown in the chart. Notice that about 30% of the students earned no points for this item, with almost half of these not attempting a response at all (nr). The distribution of scores ranged from 0 to 9 points with an overall mean of 4.2 points.

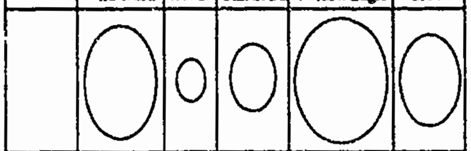
The results available from practical testing are quite different from the results of written test items. The individual practical test was designed to provide a description of the level of skills demonstrated across an entire educational system, not of one individual student. The observer notes what specific procedures were used for each student and whether or not they needed any of the hints. These data are summed across all students sampled at a given age or grade level. The overall results that are published are in terms of the percentage of students that successfully performed each skill or procedure.

The results for the "Food Preference of Mealworms" task is presented in Figure 6.19 (Harlen, Black & Johnson, 1981). The results are summarized in the same order as the items in the Observer Checklist. The observer checks a blank on the checklist each time an individual student performs a particular sub-task. For the specific skills on which boys and girls performed quite differently (differences of 5% or more), those data were also presented on this summary

Graphical Representation of Responses to a Test Item

Question Page

This table shows the eggs of 5 birds (drawn proportionally) and the usual number of days for the eggs to hatch.



	Wild Duck	Robin	Blackbird	Golden Eagle	Crow
Usual number of days to hatch	30	13	14	40	19

A magpie's egg is this size

(a) Use the information in the table to say how long the magpie's eggs are likely to take to hatch.
About______________________

(b) How did you decide it takes this time?______________________

Mark Scheme

Response	Marks
a) Allow 14 to 19 (inc)	1
Greater than 19 or less than 14	0
b) Statement mentioning that there is some connection between size and number of days to hatch, so that the magpie's egg must be between 14 and 19 days because its size is between the blackbird and the crow.	2
Statement that it is between the blackbird and crow without mention of the pattern <u>or:</u> general reference to size such as "I decided by how big the egg was" <u>or:</u> any reference to pattern in size though (a) is wrong <u>or:</u> implicit use of size	1
Irrelevant or incomprehensible	0

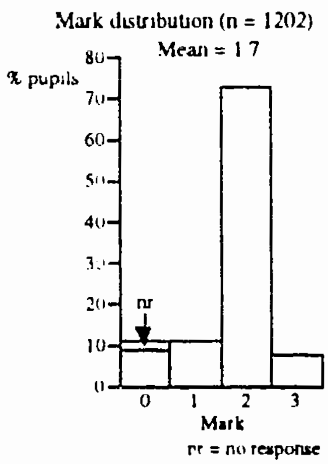


Figure 6.16. Graphical Representation of Response to a Test Item

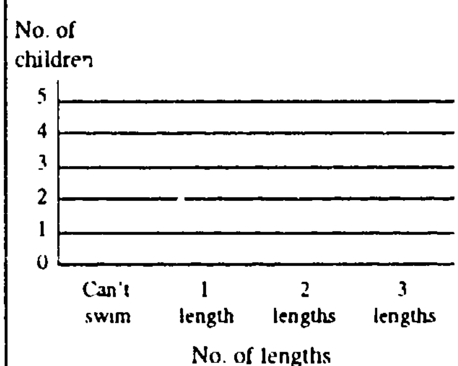
Scoring Student Responses to Tabular Data

Question page

This table shows how far each of these children can swim.

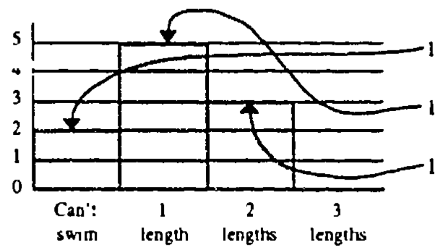
Name	Number of lengths
Mike	2
Dennis	Can't swim
Judy	2
John	1
Mary	1
Jill	2
Ian	1
Sue	1
Jane	Can't swim
Alan	1

Draw a bar chart to show how many children can swim each distance.



Mark scheme

Responses | Marks



Vertical lines correctly drawn at equal intervals (all three) 1

Maximum = 4
(If line graph drawn, give maximum of 2 if correct)
(If bar line chart drawn maximum is 3 marks)

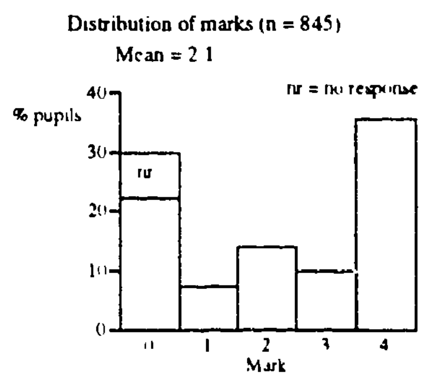


Figure 6 17 Scoring Student Responses to Tabular Data

sheet.

Just to the left of the overall percentage data is listed a letter (S, 0, C, or R) for some items. These letters correspond to four distinct modes of student interaction with the investigation. The letter "S" relates to *Strategy*—the approach the student used to attempt to answer the que·tion posed. The category includes the translation of the question into a set of procedures, such as controlling the amounts of food offered, the distances involved, etc. The "O" categor· represents the methods used to *Operationalize* the strategy. In this category are the placement of food, timing, and counting. The third category, represented by

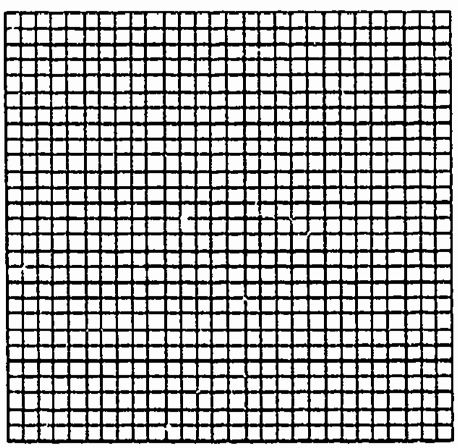
Graphical representation of Responses to a Test Item

Question page

Richard measured his bean plant every week so that he could see how fast it was growing. He started (0 weeks) when it was just 5 cm high. These were the heights for the first four weeks:

0 weeks ——	5 cm
1 week ——	15 cm
2 weeks ——	30 cm
3 weeks ——	40 cm
4 weeks ——	45 cm

Draw a graph to show how the height changed with time.



Mark Scheme

Response	Marks
One axis marked out for height	1
This axis labelled in cm	1
One axis marked out for time	1
This axis labelled in weeks	1
Height vertical, time horizontal	1
Height scale suitably chosen	1
Time scale suitably chosen	1
4 or 5 points correctly plotted	2
2 or 3 points correctly plotted	1
1 point correctly plotted	0

Maximum = 9 marks

Distribution of marks (n = 828)

Mean = 4.2

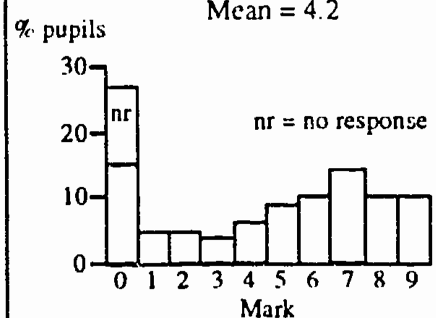


Figure 6.18. Graphical Representation of Responses to a Test Item

"C," relates to the judgment which the students are expected to make as to whether the results reported were *Consistent* with what was observed to happen. The "R" category includes the activities of *Recording* results and making notes (even if only rough or brief). The information grouped by these categories should be useful for teachers and curriculum developers to determine the effectiveness of the investigation for all aspects of the activity. Detailed descriptions of these four categories are provided in Figure 6.20 (Harlen, Black & Johnson, 1981).

An assessment of science inquiry skills was included in the NAEP Science Study conducted in 1972-73. One of the activities used with nine-year-olds was an item assessing, basically, "conservation of volume." The results were used to provide national "indicators" of levels of achievement. A sample of the results

Results of Laboratory Practical Test Item by Gender

	Student Interaction Code*	Overall	Boy/Girl Performance % B	G
Uses hand lens correctly		78	85	71
Hint given		13		
Deliberately provides mealworm with choice, i.e., at least 2 foods at once	(S)	48		
Employs an effective strategy such as: (i) uses 6 or more mealworms if all 4 foods compared at once (ii) compares foods in all possible pairs with 1 mealworm (iii) tries at least 4 mealworms with one food at a time	(S)	41	37	44
Attempts to provide equal quantities of different foods	(S)	48		
Puts approximately equal quantities of food	(O)	61		
Attempts to release mealworms at equal distance from all foods or arranges mealworm to be randomly distributed around piles of food	(S)	47		
Arranges to release mealworms from point equidistant from foods, or places meal-worms randomly around foods	(O)	45		
Arranges for all mealworms to have same time to choose (i.e., puts them all down together or uses clock)	(O)	51	49	54
Uses clock to time definitive events	(O)	16		
Allows about 4-7 minutes for mealworms to make choice (not necessarily timed)	(O)	70		
Examines behavior carefully (to see if food being eaten)	(O)	73	78	69
Counts mealworms near each pile after a certain time (or notes which food meal worm is on for strategy (ii) above)	(O)	39		
Makes notes at (a) (however brief)	(R)	58		
Records details such as time of choice and numbers near each food	(R)	18		
Can read stop clock correctly (to nearest second)		66	72	60
Makes a record of findings at (b) without prompting	(R)	72	75	70
Results at (a) and (b) consistent with evidence (even if only rough)	(C)	81	78	84
Results based on and consistent with quantitative evidence	(C)	30	28	33

* See paragraph below for explanation of student interaction codes (S, O, R, C).

Figure 6.19. Results of Laboratory Practical Test Item by Gender

Description of Student Interaction Strategies

Four main components have been defined and check-points related to them. These components are briefly defined as follows:

Component S	Check-points relating to strategy - approaching the investigation in a way which will lead to an answer of the kind required by the question posed. This includes decisions which the pupil has to make to translate the question into action rather than the details of how this action is carried out.
Component O	Check-points relating to the methods used to obtain results; attention to factors which affect accuracy in measurements or observations. It includes anything relating to the rigor with which the strategy is put into operation. Repetition of measurements or observations in cases where something has gone wrong is included hut not routine repetition (which has been seen to be very infrequent and is therefore excluded).
Component C	Check-points relating to the judgments that testers were required to make as to whether the results reported by the pupil were consistent with what was observed to happen.
Component R	Check-points relating to the use made of the spaces in the pupil's paper for making notes or recording results. It includes qualitative judgments about the adequacy of the record but not judgments about the accuracy of results which are included in component C.

Figure 6.20. Description of Student Interaction Strategies

for this item are summarized in Figure 6.21 (NAEP, 1975).

Each of the children doing the activity was observed by a trained administrator. The administrator read the questions to each student. Over half of the nine-year-olds tested thought that one of the containers had more water than the other, and two-thirds of these students thought that the tall container contained more water. When asked to show the administrator how they would find out if one of the containers had more water in it, 82% were able to demonstrate an acceptable procedure.

Although the students were not asked if the procedure they used led them to change their original hypothesis, the administrators were instructed to record any changes they observed. Of the 9-year-olds who originally said one container had more water in it, 30% changed their hypothesis after collecting pertinent data. After conducting the test, they felt the containers had equal amounts of water in them. This activity appeared to be a good learning experience for the students,

Assessment Task for Conservation of Volume

Each 9-year-old received two pre-measured containers of colored water. One was a tall container of red-colored water; the other a short container of green-colored water. Each container had 40 milliliters of water in it, but the student was not told this. The following materials were placed in front of each student: a 12-inch ruler, two clear plastic glasses, a 100-milliliter graduated cylinder, and some paper towels.

In order to get the students to form a hypothesis before attempting the activity, they were asked the following question: "Do you think one of the containers has more water in it?" The percentage of students giving each answer is shown below:

Yes	58%	I don't know	1%
No	40%	No response	0%

Approximately two out of every three of the 58% of the 9-year-olds who responded "yes" thought the amount of water in the tall, red container was greater.

The students were asked to show the administrator how they would find out if one of the containers had more water in it. They were told they could use any of the materials that had been given to them. 82% of the 9-year-olds were able to demonstrate an acceptable testing procedure in this activity. The various types of procedures used are shown in Table 1.

Table 1.
Procedures Used to Test a Hypothesis: Colored-Water Activity

Poured green water into one plastic glass and red water into another and compared them	54%
Poured the colored water from each container into the graduated cylinder in turn and compared the measurements	22%
Poured the red water into a plastic glass and compared it to the green water in the container	6%
Unacceptable procedures	14%
I don't know	2%
No response	2%

Figure 6.21 Assessment Task for Conservation of Volume

Main Findings from Science Survey on Process Skills for Il-year-olds

The surveys provide information about children's performance in the tests, about their reactions to science activities, and about the provision for science in the schools. The main results under each of these headings are summarized below.

Children's performance

MOST Il-year-olds
- set about practical investigations in a relevant manner
- observed broad similarities and differences between objects
- read the scales of simple measuring instruments correctly
- classified objects on the basis of observed properties
- read information from flow charts, tables, pie charts, and isolated points from line graphs.

ABOUT HALF Il-year-olds
- reported results consistent with the evidence from their investigations
- were more fluent at observing differences than similarities between objects
- made predictions based on observations
- suggested controis in planning parts of investigations
- used given information to make reasonable predictions
- applied science concepts to solve problems
- proposed alternative hypotheses to explain a given phenom enon
- added information to a partially completed graph or chart

FEW Il-year-olds
- repeated measurements or observations to check results
- controlled variables necessary to obtain good quantitative results
- recorded the observation of fine details of objects
- observed the correct sequence of events
- produced an adequate plan for a simple investigation
- gave good explanations of how they arrived at predictions
- described patterns in observations or data in terms of general relationships

Figure 6.22. Main Findings from Science Survey on Process Skills for 11-year-olds

as well as a test of their ability to demonstrate a simple testing procedure.

The last example in this section is the "Main Findings" from the British APU project (Harlen, 1983). These conclusions (Figure 6.22) were obtained from that project's research with youngsters age eleven. Note that the descriptions include details of skills and performances but *not* numerical results.

These findings were grouped into three categories. The descriptors within the first group - things that *most* youngsters are able to do - gives one a very precise understanding of what has been gained through their science programs. These are the skills in which the vast majority of students in almost all schools have acquired a great deal of proficiency. The second category—things that about *half* of the Il-year-olds could do—are skills that are being accomplished in some but not all elementary schools. It brings to the attention of teachers that there are some impediments to the complete mastery of these skills. It might be that some of the instructional materials or activities are inappropriate for youngsters at this age level. It might also be that too few examples are being used in some schools to allow most of the students to fully understand these concepts and to perform these skills. Further, it might be that some of the youngsters in these schools have not progressed sufficiently in terms of cognitive development to master these skills. These skills are being accomplished in some schools with some students. It might be useful to try to find out in what schools these skills are being mastered. Then one could determine what is different between those schools and settings and others in which students are not mastering these skills.

The last category summarizes those skills that only a few 11-year-olds are able to perform. It is this area that needs the major attention of teachers, as these skills are not being mastered by the student sample. Again, there may be a number of possible reasons for this performance. Assuming the skills are appropriate for this age range of students, several factors of the instructional system need to be considered. One should realize that the entire thrust of this evaluation is a description of the program and the instruction, rather than the students.

Grading and Evaluating Students

Once a set of information has been collected, a variety of conclusions can be formulated. Dramatically different kinds of statements result from norm-referenced and criterion-referenced systems.

In a criterion-referenced system, each student's performance would be judged against some previously established standard or criterion. Each student will be described as being "successful" or as having "mastered" a particular unit of instruction once that student demonstrates the skill or knowledge objectives. A basic need of this system is a clear, understandable description of the content or skill outcome.

A norm-referenced system compares each student with the performance of some norming group (class, school, state, or national groups). A common way of combining information for grading purposes is a "formula" which weights various information, such as test scores, projects, and reports. The formula method of combining data can vary widely, but the "bottom line" is that a letter, number, or descriptive evaluation is the result. This will describe a student's rank or position with respect to the other students of the norming group. One possible formula for determining grades of students in a science experience would be the following:

Written tests	30%
Practical tests	30%
Observation	20%
Projects	10%

Reports	10%

This example is *hypothetical*! It represents a situation in which the teacher(s) believe that written and practical tests are the most important ways of assessing performance in science. Each is weighted 30% of the total. Data collected by a variety of observations, perhaps some during problem-solving situations, is also important, accounting for 20% of a student's grade. Information collected from the analysis of projects and of reports is less valued, each contributing 10% to the final evaluation. This kind of system is usually applied to data obtained during some marking period, such as the ten-week grading period, a semester, or the entire school year. The amount of information that has been collected can vary widely, as can the nature of the assessment procedures used in each of the contributing areas (written tests, projects, etc.). It is considered essential to tap the different skills or channels through which students can demonstrate what they have learned in science. The "formula" becomes a way to combine these separate inputs into a singular evaluation.

An example of the criterion-referenced system will be described. This example is based on the kinds of reports common to most elementary schools and a scheme of process mastery (Harlen, 1985). Rather than giving a single grade or assessment to a student for the science area, this example uses four categories of science performance. The four major categories are:

1. Knowing science information
2. Using science concepts and generalizations
3. Doing written reports and projects
4. Experimenting or investigating

For each of the categories, we are proposing that students be described as being at one of the levels illustrated below. The specific wording of each descriptor must be tailored to the specific skills that students at a given grade level are expected to do.

Throughout this section we have stressed the need to assess students via a variety of formats, which would include written tests, observations, practical tests, etc. Based on information from these various modes, one could have a wealth of information for use in describing the level of proficiency which a given student has in these categories. For each category we have listed three statements describing levels of performance, from a rather weak stage of development to a robust mastery at the highest level. It is assumed that most students' performance could be expressed in one of these three levels. While a few students might be described at the same level in all four skill areas, it is more likely that students will be at different levels for these different skills.

Knowing Science Information

1. Responds only in terms of specific examples experienced in class or presented in instructional materials.

2. Responds in terms of generalizations of these experiences but is unable to Show relationships or to go beyond that which was experienced.

3. Demonstrates thorough understanding by applying information in a new context or by explaining relationships, implications, or consequences.

Using Science Concepts and Generalizations

1. Rarely connects previous learning with new situations in which it could be applied unless told what skill or idea is relevant.

2. Uses previous experiences in new situations once the relationship between the new and previous situation has been pointed out.

3. Works out what earlier learning could be applied in a new context by using relationships between one situation and another.

Doing Written Reports and Projects

1. What he writes or says is disorganized and difficult to follow; takes time to understand information in books or verbal directions.

2. Seems to have a clear idea of what he wants to express but does not always find the words to put it precisely or concisely; prefers to seek information orally than to use books.

3. Expresses himself clearly, using words appropriately and economically and at a level which can be understood by whomever receives the message; expands his knowledge through reading.

Experimenting/Investigating

1. Is unable to progress from one point to another in a practical investigation or inquiry without help, failing to grasp the overall plan.

2. Tries things out somewhat unsystematically unless the various steps in a practical inquiry are planned out for him, in which case he uses materials and collects results satisfactorily.

3. Has a clear idea of the reason for the various steps in an investigation; can work through them systematically, making reasonable decisions with only occasional guidance.

A basic concern of every grading system, whether norm- or criterion-referenced, is the efficient storage of the information which is carefully collected. Most teachers have evolved or learned a good system for storing data from tests and quizzes. The information as to performance levels of various skills might present a new difficulty. Some of this information can be gathered in group or practical tests, but much can be gathered from performance and behaviors observed or reported from various class activities and investigations. To assess every skill with every student on each report and through each investigation becomes an incredibly massive task. The assessment card illustrated in Figure 6.23 was designed to help teachers compile measures of student performance on various skills via a sampling system (Lunetta, Hofstein & Giddings, 1981). Although it would be ideal to provide a detailed assessment of each science skill in each activity or report, the time demands are prohibitive. The system reflected in the card suggests that for each lab activity the teacher chooses a few skills

Laboratory Skills Assessment

Name: Geoffrey Giddings Course: Biology

Experiment/ Unit Assessed	Date	A. Planning & Design	B. Manipulative Skills & Conduct of Experiment	C. Observations & Recording Data	D. Interpretations of Data & the Experiment	E. Responsibility/ Initiative/ Work Habits
1	9/25/80		5	4		
2	10/25/80	3	7			2
3	10/27/80		3	4		
4	11/15/80		7			1
5	12/02/80	4		4	6	8
6	12/13/80		5			
7	1/06/81				4	
8	1/24/81	7	8	6		2
9	2/03/81			5	6	3
10	2/17/81		7			5

Figure 6.23. Laboratory Skills Assessment Chart

appropriate to that particular task on which to make specific observation or assessment. As these are collected over a period of time, the overall assessment obtained is balanced and meaningful.

Program Evaluation

Having illustrated several methods for gathering data about the science achievement and performance of elementary school students, the question arises as how to combine this information for a comprehensive assessment program. One must keep in mind the amount of testing which is expected of each student, so as to not cause test anxiety. Further, some of the assessment procedures require extensive teacher and class time.

One technique used to organize the administration of the wide variety of assessment items and to minimize individual student testing burden is called "matrix" sampling. The following chart (Figure 6.24) is an illustration of how matrix sampling could be used in the assessment of elementary science programs.

Along the left-hand margin is a list of student names organized alphabetically or by school and class records. This chart would likely include a given class or the students in one grade. This student list is necessary so one can allocate randomly different items or sub-tests to individual students. The aim of this procedure is to obtain a balanced and accurate perspective of the performance of the class or grade, not of the individual student.

The written test section of the illustrative matrix is composed of five sub-tests. As each student responds to just 1/5 of the items in the total pool, one can obtain "cove ıge" of 5 times as many test items than if each student received all the test items If a pool of 100 written test items were constructed, each child would be expected to answer just 20 of those items. The 20 items assigned to each sub-test would need to be chosen so as to represent the content and objectives in as balanced a way as possible. In addition to the "content" items within the written test, several categories of the science processes may be validly assessed by the written test format.

Other science performance categories are better assessed with group or individual practical tests. These are more time-consuming to prepare and administer but are the most appropriate way to assess these skills. If the skills are

Program Assessment Chart

	Written Tests					Group Practical		Individual Practical				
Students	I	II	III	IV	V	A	B	1	2	3	4	5
A	x					x		x				
B		x					x		x			
C			x			x				x		
D				x			x				x	
E					x	x						x
F	x						x	x				
G		x				x			x			
H			x				x			x		
I				x		x					x	
J					x		x					x
K	x					x		x				
L		x					x		x			
M			x			x				x		
N				x			x				x	
O					x	x						x

Figure 6.24. Program Assessment Chart

not assessed, students, administrators, teachers, and parents come to believe that such skills "really are not important." The matrix in Figure 6.24 demonstrates one way to efficiently administer samples of practical tests to sub-groups of students.

In the example cited . .ere are two forms of the group practical test. Two forms of practical tests (e.g., Form A and B) could be composed of different tasks or situations in which students are expected to perform certain science skills. The recommended administration is that each student respond to just one of these practical test forms (A or B).

A number of educators believe that some science skills can only be assessed by the detailed report from a trained observer. Such skills as the use of apparatus and measuring instruments and the performance of investigations are examples of such skills. How many tasks can be administered in this manner depends on the length of the tasks and the time available for the observer(s). In the matrix we have illustrated a case in which each student completes only one of five individual tasks in the presence of an observer. This system could be adapted to a different number of available individual tests. However, it is recommended to administer initially only one individual practical task to each student. If the interest and respect for such assessment grows, one could expand the number of tasks used with each student.

In summary, the sampling matrix illustrates one system for obtaining student responses from a large pool of written and practical tests WITHOUT over-burdening each individual student. As a matter of fact, each student would respond to 1/4 of the total number of sub-tests or tasks in the pool. Each would respond to one of five sub-tests of the written test, one of two group practical test forms, and one of five individual practical tests for a total of three out of twelve possible sub-tests. This system has been recommended for several reasons; the main reason, though, is to accomplish the goal of program evaluation without excessive burden on the students. If a teacher or school is interested in assessing individual students, additional testing sessions could be planned which would incorporate some of the remaining testing situations.

References

Addison Wesley Publishing Co. (1979). *Sequential test of educational progress, level E.* Menlo Park, California: Author.

Addison Wesley Publishing Co. (1979). *Sequential test of educational progress, level G.* Menlo Park, California: Author.

Aho, W., et.al. (1974). *The McGraw-Hill evaluation program for CSS.* Webster, Missouri: McGraw-Hill.

American Association for the Advancement of Science. (1968). *Science, a process approach.* Lexington, MA: Ginn.

Bloom, B.; Engelhart, M.; Furst, E.; Hill, W.; and Krathwohl, D. (Eds.). (1956). *Taxonomy of educational objectives handbook I: cognitive domain.* New York: McKay.

Bryce, T. and Robertson, T. (1985). "What can they do? A review of practical assessment in science." *Studies in Science Education*, 12: 1-24.

California Test Bureau/McGraw Hill. (1973). *Comprehensive test of basic skills, level C, form S.* Monterrey, California: Author.

Chiarello, M. (1989). Unpublished materials. Manhasset School District, Manhasset, New York: Author.

Frayer, D.; Frederick, W.; and Klausmeir, H. (1969). *A scheme for testing the level of concept mastery.* (Working paper 16). Madison, Wisconsin: Wisconsin Research and Development Center for Cognitive Learning.

Griffith-Miles, M. (1989). "Which food does Fluffy like best?" Manhasset, NY: personal communication.

Harcourt, Brace, and Jovanovich. (1973). *Stanford achievement test, primary level II.* New York: Author.

Harcourt, Brace, and Jovanovich. (1978). *Metropolitan achievement test, intermediate level, form JS.* New York: Author.

Harlen, W. (1983). *Science at age 11: science report for teachers 1.* London: Assessment of Performance Unit (Department of Education and Science).

Harlen, W. (1985). *Primary science: taking the plunge.* London: Heineman.

Harlen, W.; Black, P.; and Johnson, S. (1981). *Science in schools, age 11: report no. 1.* London: Assessment of Performance Unit (Department of Education and Science).

Harlen, W.; Darwin S.; and Murphy, M. (1978). *Match and mismatch, raising questions: leader guide.* Edinburgh, Scotland: Oliver and Boyd.

Harlen, W.; Palacio, D.; and Russell, T. (1984). *Science assessment framework, age 11: science report for teachers 4.* London: Assessment of Performance Unit (Department of Education and Science).

Instructional Objectives Exchange. (1972). *Life sciences, K-6.* Los Angeles: Author.

Jacobson, W. and Doran, R. (1985). "Second International Science Study: some preliminary results." *Phi Delta Kappan*, February, 1985.

Jefferson County Schools. (1982). *Elementary science test.* Lakewood, Colorado: Author.

Lunetta, V.; Hofstein, A.; and Giddings, G. (1981). "Evaluating science laboratory skills." *The Science Teacher*, 48(1): 22-25.

Lunetta, V. and Tamir, P. (1979). "Matching lab activities with teaching goals." *The Science Teacher*, 46 (5): 22-24.

Maki, A. (1985). *Student Report.* Manhasset, NY: Aim Project.

Manhasset Public Schools. (1985). *Science in the elementary school: a statement of policy*. Manhasset, New York: Author. (Unpublished document). Written by E. Meng.

Meng, E. and Doran, R. (1990). "What research says about appropriate methods of assessment." *Science and Children, 28*(1): 42-45.

Michigan Educational Assessment Program. (1985). *Science test, grade 4, form C*. Lansing, Michigan: Author.

National Assessment of Educational Programs. (1975). *Selected results from the National Assessment of Sciences: selected principles and procedures*. (Report no. 04-S-42). Princeton, NJ: Author.

National Assessment of Educational Programs. (1987). *Learning by doing*. Princeton: Educational Testing Service.

New York State Educational Department. (1985). *Elementary science syllabus*. Albany, New York: Author.

Osborne, R. and Freyberg, P. (1985) *Learning in science*. London: Heineman.

Rand McNally. (1978). *SCIIS evaluation supplements*. Chicago: Author.

Science Olympiad. (1986). *Aerodynamics activity*. Rochester, NY: Author.

Second International Science Study. (1984). *Science test, grade 5*. New York: Teachers College, Columbia University.

Second International Science Study. (1986). *Science process test - grade 5*. New York: Teachers College, Columbia University.

Shayer, M. and Adey, P. (1981). *Toward a science of science teaching*. London: Heineman.

Voelker, A. and Sorenson, J. (1971). *Classification and analysis of science concepts.* (Working page 58). Madison, Wisconsin: Wisconsin Research and Development Center for Cognitive Learning.

Wisconsin Department of Public Instruction. (1970). *A guide to science curriculum development.* (Bulletin no. 161). Madison, Wisconsin: Author.

A Brief Guide to ERIC

The Educational Resources Information Center

Office of Educational Research and Improvement
U.S. Department of Education

What is ERIC?

The Educational Resources Information Center (ERIC) is a national education information network designed to provide users with ready access to an extensive body of education-related literature. Established in 1966, ERIC is supported by the U.S. Department of Education, Office of Educational Research and Improvement.

The ERIC database, the world's largest source of education information, contains over 735,000 abstracts of documents and journal articles on education research and practice. This information is available at more than 2,800 libraries and other locations worldwide.

You can access the ERIC database by using the print indexes *Resources in Education* and *Current Index to Journals in Education*, online search services, or CD-ROM at many libraries and information centers. The database is updated monthly (quarterly on CD-ROM).

The ERIC System

The ERIC System, through its 16 subject-specific Clearinghouses, 4 Adjunct Clearinghouses, and four support components, provides a variety of services and products that can help you stay up to date on a broad range of education-related issues. Products include research summaries, publications on topics of high interest, newsletters, and bibliographies. ERIC system services include computer search services, reference and referral services, and document reproduction. ACCESS ERIC, with its toll-free number, 1-800-LET-ERIC, informs callers of the services and products offered by ERIC components and other education information service providers.

ERIC Reference and Referral Services

With the world's largest educational database as a resource, ERIC staff can help you find answers to education-related questions, refer you to appropriate information sources, and provide relevant publications. ERIC components answer more than 100,000 inquiries each year. Questions should be directed to ACCESS ERIC or a specific Clearinghouse.

Specific documents: Requests for documents in the ERIC database for which you have an accession number (ED number) should be referred to an information provider near you. Call ACCESS ERIC to locate the nearest ERIC education information provider.

Subject-specific topics: Subject-related questions should be directed to the particular ERIC Clearinghouse whose scope is most closely associated with the subject matter involved. Or, call ACCESS ERIC for a referral.

Computer searches: Requests for a computer search should be directed to one of the search services listed in the *Directory of ERIC Information Service Providers,* available from ACCESS ERIC.

ERIC Clearinghouse publications: Requests for a publication produced by an ERIC Clearinghouse should be directed to the specific Clearinghouse.

Major ERIC Products

ERIC produces many products to help you access and use the information in the ERIC database:

Abstract Journals: ERIC produces two monthly abstract journals. *Resources in Education* (RIE), a publication announcing recent education-related documents, and the *Current Index to Journals in Education* (CIJE), a periodical announcing education-related journal articles, is available through Oryx Press (1-800-457-6799). Many libraries and information centers subscribe to both monthly journals.

***All About ERIC*:** This guide provides detailed information on ERIC, its products and services, and how to use them. Free copies are available from ACCESS ERIC.

***Catalog of ERIC Clearinghouse Publications*:** The *Catalog* lists publications produced by the ERIC Clearinghouses and support components, prices, and ordering information. Free copies of the *Catalog* are available from ACCESS ERIC.

***The ERIC Review*:** This journal discusses important ERIC and education-related developments. For a copy, call ACCESS ERIC.

Information Analysis Products: ERIC Clearinghouses produce reports, interpretive summaries, syntheses, digests, and other publications, many free or for a minimal fee. Contact the Clearinghouse most closely associated with your interests for its publications list. Call ACCESS ERIC for a free copy of the *Catalog of ERIC Clearinghouse Publications.*

Microfiche: The full text of most ERIC documents is available on microfiche. Individual documents and back collections on microfiche are available. Call the ERIC Reproduction Document Service (EDRS) for more information.

Thesaurus of ERIC Descriptors - The complete list of index terms used by the ERIC System, with a complete cross-reference structure and rotated and hierarchical displays, is available from Oryx Press.

ERICTAPES - Computer tapes of the ERIC database are available by subscription or on demand from the ERIC Facility (write for a price list).

ERIC Document Delivery

Documents: EDRS is the primary source for obtaining microfiche or paper copies of materials from the ERIC database. EDRS can provide full-text copies of most documents announced in Resources in Education (RIE), and ERIC's microfiche collection is available by monthly subscription from EDRS. EDRS also sells microfiche and paper copies of individual documents on request. For more information, call EDRS at (800) 443-ERIC.

Journal Articles: Two agencies that provide reprint services of most journal articles announced in *Current Index to Journals in Education* (CIJE) are listed below. Some journals do not permit reprints; consult your local university or local library to locate a journal issue. Or, write directly to the publisher. Addresses are listed in the front of each CIJE.

University Microfilms International (UMI)
Article Clearinghouse
300 North Zeeb Road
Ann Arbor, MI 48106
Telephone: (800) 732-0616

Institute for Scientific Information (ISI)
Genuine Article Service
3501 Market Street
Philadelphia, PA 19104
Telephone: (800) 523-1850

ERIC Information Retrieval Services

The ERIC database is one of the most widely used bibliographic databases in the world. Last year, users from 90 different countries performed nearly half a million searches of the database. The ERIC database currently can be searched via four major online and CD-ROM vendors (listed below). Anyone wishing to search ERIC online needs a computer or terminal that can link by telephone to the vendor's computer, communications software, and an account with one or more vendors.

The *Directory of ERIC Information Providers* lists the address, telephone number, and ERIC collection status for more than 900 agencies that perform searches. To order a copy, call ACCESS ERIC (1-800-LET-ERIC).

Online Vendors

BRS Information Technologies
8000 Westpark Drive
McLean, VA 22102
Telephone: (703) 442-0900
(800) 289-4277

Dialog Information Services
3460 Hillview Avenue
Palo Alto, CA 94304
Telephone: (415) 858-2700
(800) 334-2564

OCLC (Online Computer Library Center, Inc.)
6565 Frantz Road
Dublin, OH 43017-0702
Telephone (614) 764-6000
(800) 848-5878
(Ext. 6287)

CD-ROM Vendors

Dialog Information Services (same address as above)

Silver Platter Information Services
One Newton Executive Park
Newton Lower Falls, MA 02162-1449
Telephone: (617) 969-2332
(800) 343-0064

ERIC Components

Federal Sponsor

Educational Resources Information Center (ERIC)
U.S. Department of Education
Office of Educational Research and Improvement (OERI)
555 New Jersey Avenue N.W.
Washington, DC 20208-5720
Telephone (202) 219-2289
Fax: (202) 219-1817

Clearinghouses

Adult, Career, and Vocational Education
The Ohio State University
1900 Kenny Road
Columbus, OH 43210-1090
Telephone: (614) 292-4353; (800) 848-4815
Fax (614) 292-1260

Counseling and Personnel Services
University of Michigan
School of Education, Room 2108
610 East University Street
Ann Arbor, MI 48109-1259
Telephone: (313) 764-9492
Fax: (313) 747-2425

Educational Management
University of Oregon
1787 Agate Street
Eugene, OR 97403-5207
Telephone: (503) 346-5043
Fax: (503) 346-5890

Elementary and Early Childhood Education
University of Illinois
College of Education
805 W. Pennsylvania
Urbana, IL 61801-4897
Telephone: (217) 333-1386
Fax (217) 333-5847

Handicapped and Gifted Children
Council for Exceptional Children
1920 Association Drive
Reston, VA 22091-1589
Telephone: (703) 620-3660
Fax: (703) 264-9494

Higher Education
The George Washington University
One Dupont Circle N.W., Suite 630
Washington, DC 20036-1183
Telephone: (202) 296-2597
Fax: (202) 296-8379

Information Resources
Syracuse University
Huntington Hall, Room 030
Syracuse, NY 13244-2340
Telephone: (315) 443-3640
Fax: (315) 443-5732

Junior Colleges
University of California at Los Angeles (UCLA)
Math-Sciences Building, Room 8118
405 Hilgard Avenue
Los Angeles, CA 90024-1564
Telephone: (213) 825-3931
Fax: (213) 206-8095

Languages and Linguistics*
Center for Applied Linguistics
1118 22nd Street N.W.
Washington, DC 20037-0037
Telephone: (202) 429-9551 and (202) 429-9292
Fax: (202) 429-9766 and (202) 659-5641
**Includes Adjunct ERIC Clearinghouse on Literacy Education for Limited English Proficient Adults*

Reading and Communication Skills
Indiana University
Smith Research Center, Suite 150
2805 East 10th Street
Bloomington, IN 47408-2698
Telephone: (812) 855-5847
Fax: (812) 855-7901

Rural Education and Small Schools
Appalachia Educational Laboratory
1031 Quarrier Street
P.O. Box 1348
Charleston, WV 25325-1348
Telephone: (800) 624-9120 (outside WV)
(800) 344-6646 (inside WV)
(304) 347-0400 (Charleston area)
Fax: (304) 347-0487

Science, Mathematics, and Environmental Education
The Ohio State University
1929 Kenny Road
Columbus, OH 43210-1080
Telephone: (614) 292-6717
Fax (614) 292-0263

Social Studies/Social Studies Education**
Indiana University
Social Studies Development Center
2805 East 10th Street, Suite 120
Bloomington, IN 47408-2373
Telephone: (812) 855-3838
Fax: (812) 855-7901
***Includes Adjunct ERIC Clearinghouse on Art Education; and the National Clearinghouse for U. S.-Japan Studies*

Teacher Education
American Association of Colleges for Teacher Education (AACTE)
One Dupont Circle N.W., Suite 610
Washington, DC 20036-2412
Telephone: (202) 293-2450
Fax: (202) 457-8095

Tests, Measurement, and Evaluation
American Institutes for Research (AIR)
Washington Research Center
3333 K Street N.W.
Washington, DC 20007-3893
Telephone: (202) 342-5060
Fax: (202) 342-5033

Urban Education
Teachers College, Columbia University
Institute for Urban and Minority Education
Main Hall, Room 300, Box 40
525 West 120th Street
New York, NY 10027-9998
Telephone: (212) 678-3433; Fax (212) 678-4048

Adjunct Clearinghouses

National Clearinghouse on Literacy Education
Center for Applied Linguistics
1118 22nd Street N.W.
Washington, DC 20037
Telephone: (202) 429-9292
Fax: (202) 659-5641

Adjunct ERIC Clearinghouse for Art Education
Indiana University
Social Studies Development Center
2805 East 10th Street, Suite 120
Bloomington, IN 47405-2698
Telephone: (812) 855-3838
Fax: (812) 855-7901

National Clearinghouse for U. S.-Studies
Indiana University
Social Studies Development Center
2805 East 10th Street, Suite 120
Bloomington, IN 47408-2698
Fax: (812) 855-7901

Adjunct ERIC Clearinghouse on Chapter 1
Chapter 1 Technical Assistance Center
PRC, Inc.
2601 Fortune Circle East
One Park Fletcher Building, Suite 300-A
Indianapolis, IN 46241
Telephone: (317) 244-8160, (800) 456-2380
Fax: (317) 244-7386

Support Components

ACCESS ERIC
Aspen Systems Corporation
1600 Research Boulevard
Rockville, MD 20850-3166
Telephone: (800) LET-ERIC
Fax: (301) 251-5212

ERIC Document Reproduction Service
7420 Fullerton Road, Suite 110
Springfield, VA 22153-2852
Telephone: (301) 258-5500; (800) 443-ERIC
Fax: (301) 948-3695

Oryx Press
4041 North Central Ave., Suite 700
Phoenix, AZ 85012-3399
Telephone: (602) 265-2651; (800) 279-ORYX
Fax: (602) 26. -6250; (800) 279-4663

How to Submit Documents to ERIC

ERIC collects a variety of materials on education-related topics. Examples of materials included in the database:

Research reports
Instructional materials
Monographs
Teaching Guides
Speeches and presentations
Manuals and handbooks
Opinion papers

Submissions can be sent to the Acquisitions Department of the ERIC Clearinghouse most closely related to the subject of the paper submitted, or sent to the ERIC Processing Facility.

Notes

ERIC/CSMEE

1929 Kenny Road, Columbus, OH 43210-1080