DOCUMENT RESUME

ED 196 733 SE 034 039

AUTHOR Doran, Rodney L.
TITLE Basic Measurement and Evaluation of Science Instruction.
INSTITUTION National Science Teachers Association, Washington, D.C.
REPORT NO ISBN-0-87355-016-1
PUB DATE 80
NOTE 137p.: Not available in hard copy due to copyright restrictions. Pages 23-24, 56, 77, 81, 85, 118, and 121 removed due to copyright restrictions.
AVAILABLE FROM National Science Teachers Association, 1742 Connecticut Ave., N.W., Washington, DC 20009 (Stock No. 471-14764; no price quoted).

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Educational Assessment; Educational Objectives; Elementary Secondary Education; Evaluation; *Evaluation Methods; Grading; Higher Education; *Measurement; Program Evaluation; *Science Education; Science Instruction; Student Evaluation

ABSTRACT

Designed to be used by preservice and in-service science teachers interested in assessing the outcomes of school science programs, this publication is aimed at helping teachers do a better job of developing tests and inventories specifically for their instructional programs and students. Material is presented in six chapters entitled: (1) Trends in Measurement and Evaluation of Science Instruction; (2) Assessing Cognitive Outcomes in Science; (3) Assessing Affective Outcomes in Science; (4) Assessing the Outcomes of Science Laboratory Activity; (5) Item and Test Analysis; and (6) Grading Students in Science. Also included is a selected references section containing 97 titles. (PB)

# BASIC MEASUREMENT and EVALUATION OF SCIENCE INSTRUCTION

Rodney L. Doran
State University of New York at Buffalo

National Science Teachers Association
1742 Connecticut Avenue, NW
Washington, DC 20009

Designed and edited by
Jennifer Knerr



NSTA Stock Number
471-14764

ISBN
0-87355-016-1

# PREFACE

The materials and examples within this book were synthesized as an aid to teaching the course, "Measurement and Evaluation of Science Instruction." The course was taught in response to a perceived need for a specific focus on the unique demands of assessing the outcomes of school science programs. Although many excellent examinations exist for monitoring state or national levels of achievement, most teachers develop tests and inventories specific to the particular instructional programs they have prepared for their students. This book is intended to help teachers with that development.

The first chapter describes the changes that have occurred and that continue to occur in the objectives and strategies of science instruction. "Change" is a concept which characterizes the field of science, especially the last several decades of science education. This book focuses on the item writing, test analysis, and grading methodology necessary to keep science teachers and supervisors and their assessment and evaluation techniques "in step" with the rapidly shifting outcomes of school science programs.

Despite the fact that "learning occurs holistically," it is helpful to focus separately on the measurement of behaviors from the cognitive, affective, and laboratory domains. The relative emphasis of these domains will vary widely with the nature of each class and its students. Nevertheless, the overall evaluation plan for every science class should contain some elements from each of these three domains. Each domain may be individually conceptualized for the purpose of planning and designing appropriate data collecting techniques. Some item formats are useful for all three domains, but others are primarily useful for one domain alone. With this kind of consideration in mind, separate chapters address assessment techniques for the cognitive, affective, and laboratory domains.

Considering the massive impact on students that scores from our tests have, we have a responsibility to make our tests as valid and reliable as possible. Chapter 5 addresses this responsibility and focuses on the individual items which compose a test or inventory. Techniques for quantitatively and qualitatively analyzing each part of an item are discussed and illustrated.

The focus of the last chapter is the utilization of the data collected from tests and inventories to comprehensively and consistently monitor and report on the achievement of students. The grades and evaluations received by students are of the utmost importance to themselves, other teachers, parents, administrators, college admission people and potential employers. Several techniques and guidelines for reporting the results of evaluation efforts are described and illustrated.

Evaluation is an integal part of instruction, and the teacher is the key to all classroom learning--before, during, and after the test. This book attempts to suggest techniques that are both relevant and useful to science teachers who wish to enhance their competencies in this dimension of science teaching.

Grateful acknowledgement is extended to several individuals and companies who allowed various items and inventories to be reprinted here. These materials add much to the successful implementation of the suggestions included within. Credit is given to each at appropriate places within the text or in the list of Selected References.

I wish to extend my appreciation to Dr. John M. Fowler, who provided the initial NSTA interest in this publication, and to Dr. Helenmarie Hofman, who facilitated the successful transition of the many stages of publication and whose contributions to the editing process were significant. I wish to thank all the students who commented on early versions of this work and who encouraged completion of the task. Thanks also are due Brenda McClintock, the NSTA staff member who was responsible for the word processing of the manuscript, and Jennifer Kherr, the general editor and production manager of the project. Her questions, comments, and editorial skills have greatly enhanced the coherence and utility of the book.

Finally, of course, the responsibility for any errors that may appear rests solely with the author.

RLD

# CONTENTS

# CHAPTER ONE

# Trends in Measurement and Evaluation of Science Instruction

**Introduction**

What we teach and how we teach it: these things are changing continuously in every discipline and at every level, making teaching an exercise in the "adapt or die" regimen of pedagogical evolution. Perhaps nowhere is this dynamism so apparent as in the teaching of science, where technology serves as both product and process, coupling the considerations of content and instructional mode more closely than in any other discipline.

Such rapid currents of change create a challenge to continuity as multifaceted as the changes themselves. The last decade has urged upon us science instruction that is humanistic, individualized, value-oriented, societally-related, as well as future-focused. These shifting--and sometimes seemingly conflicting--goals of science teaching require ways of evaluating that are both fluid and functional, comprehensive and yet precise.

Diagnostic testing, criterion-referenced measurement, and minimum competency examination are but a few examples of the new forms of evaluation being proposed. Changes in the techniques of evaluation have historically lagged behind curricular and instructional innovations. Just as the lag time between a new scientific theory and its technical application is shortening, however, so science teachers are pressed to respond to rapidly shifting instructional priorities with similarly paced adaptations of evaluation techniques and instruments.

The role of "teacher as evaluator" has, in the past, assumed a priority lower than that of other roles in which science educators are cast--roles of scientist, laboratory director, curriculum planner, career counselor, and disciplinarian. Familiarity and the perception of success are keys to the setting of role priorities. Most teachers have received little formal training in evaluation techniques, and the instruction they have received has often been cluttered with confusing definitions and formulas: long on theory but short on practical application. Perhaps because of this less than ideal preparation, evaluation has traditionally tended to be formalized and concentrated in a few, isolated days of scattered quizzes and end-of-term tests, thus casting it outside the mainstream of everyday classroom activities like laboratory demonstrations or lesson planning and presentation.

The net effect, of course, has been the estrangement of the teaching and evaluation processes and, perhaps worse, the alienation of the evaluator from those being evaluated. Teachers have, understandably, found it difficult to derive satisfaction from a role for which they feel ill-prepared and in which they are perceived as educational executioners.

It doesn't have to be this way! Evaluation is a mainstream educational tool which is most valuable and least obtrusive when integrated with all phases of the instructional process.

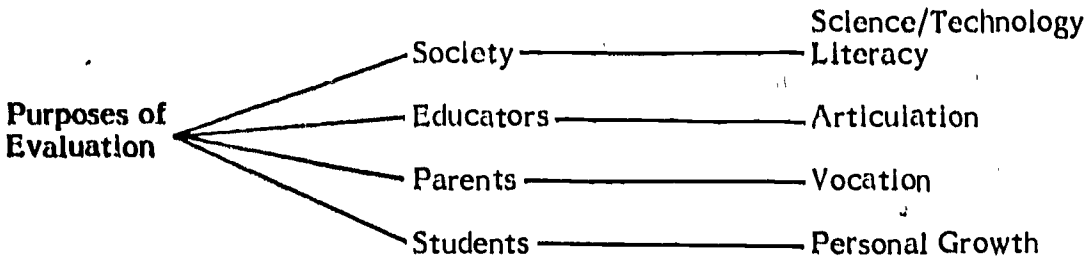
"Keeping tabs on" the students' development needn't imply "keeping under wraps" their learning behavior; a less formal, more innovative approach to evaluation can generate more creative teaching and learning, while promoting students' involvement in their own educational objectives and outcomes. This involvement, in turn, can remove some of the onus and burden of evaluation from the teacher's shoulders.

Perhaps most liberating, however, are evaluative techniques which flex with the situation at hand and with which teachers are both comfortable and conversant. Collected here are some ways of evaluating that, above all, can be readily understood, adapted, and introduced into the classroom. Some are old, some are new; some require an understanding of their theoretical underpinnings, while others create understanding through direct application. All are presented under the assumption that measurement and evaluation are basic to effective science instruction.

Here, then, are the basics of measurement and evaluation of science instruction.

## Outline of the Measurement and Evaluation Domain

To discuss specific trends and particular evaluation techniques, a general understanding of the domain is essential. The following series of figures attempts to present the many aspects and goals of evaluation and their interrelationships. The details on the figures are intended to be illustrative only and should not be interpreted as an exhaustive compilation.

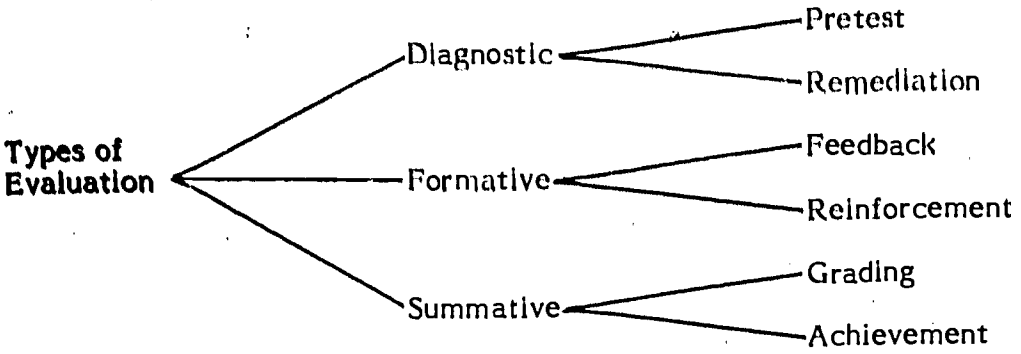



The criteria for a given evaluation program influence the kinds of data to be collected and determine the standards by which the data will be judged. Broad and timely participation in establishing these criteria is imperative. Recommendations from various parts of society, as well as from educators, parents, students, and other interested parties, should be obtained. Each of these groups may suggest a criterion of specific concern; only a few of these possible concerns are listed in the figure. Methods for obtaining recommendations from these groups will vary from open forums or committee meetings to various forms of questionnaires, checklists,and other written formats. Without a shared understanding of "why we're evaluating," the exercise will likely be futile.

For several decades now, the need has been expressed for citizens informed of the impact, procedures and limitations of the scientific and technological enterprise so predominant in 20th century America. These criteria are brought to bear in the general or liberal education of "typical" American citizens, many of whom can be aptly described as being "nonscience oriented."

We as educators are deeply concerned about how the science programs at each grade and level (elementary, middle/junior and senior) "fit together." This articulation is also important at the individual student level, for students need to possess certain understandings and skills in order to be able to learn from later science experiences.

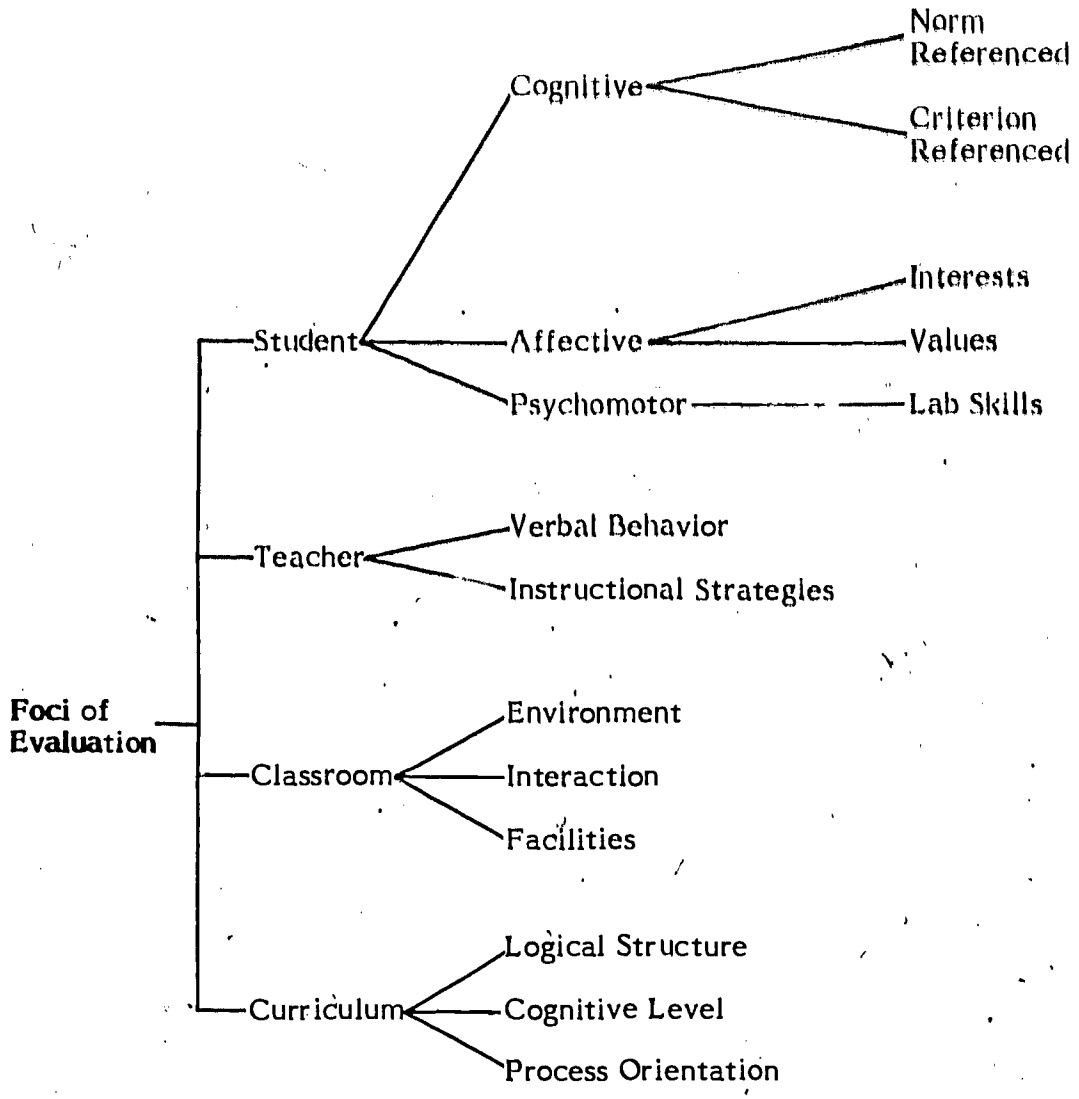
Parents often express concerns about how a specific course will help their child get a job, prepare for possible careers, or gain admission to a particular training program. At higher levels of schooling, students share this interest in vocational preparation. At lower levels, however, many students are more interested in knowing more about themselves and their "near-environment." This curiosity and "ego-centered" personal concern can be used beneficially by science teachers.




The three primary types of evaluation--diagnostic, formative, and summative--are differentiated primarily by their chronological relationship to the instructional sequence. Diagnostic evaluation normally precedes the instruction, but may be used during instruction when student learning problems arise. The results of diagnostic evaluation can provide valuable information to teachers about the knowledge, attitudes, and skills of incoming students. Such information could be the basis for individual remedial work or specific instructional arrangements. And, based on a recent review of research, Okey (69) concluded that frequent diagnostic testing can raise achievement scores.

Formative evaluation efforts are usually conducted and completed during the instructional period to provide reinforcement for student learning and feedback to the teacher for assessing progress and effectiveness. Formative evaluation is a major component of the development of science curricula by funded projects. In the classroom, too, most teachers are continually modifying their instructional package in at least minor ways, and the collection of formative data can help monitor and direct such curricular improvements.

The third kind of evaluation, the summative, is the most common. The most familiar forms of summative evaluation results are student grades and reports of achievement on completed units or courses of instruction.




Evaluation efforts can be described according to whether they focus primarily on students, teachers, classrooms, or curricula/instructional programs. The large number of specific examples provided for student evaluation does not mean that that focus of evaluation is necessarily most important. Rather, it indicates that more examples of this type have been identified and are more often discussed and applied than are evaluation efforts in the other categories.

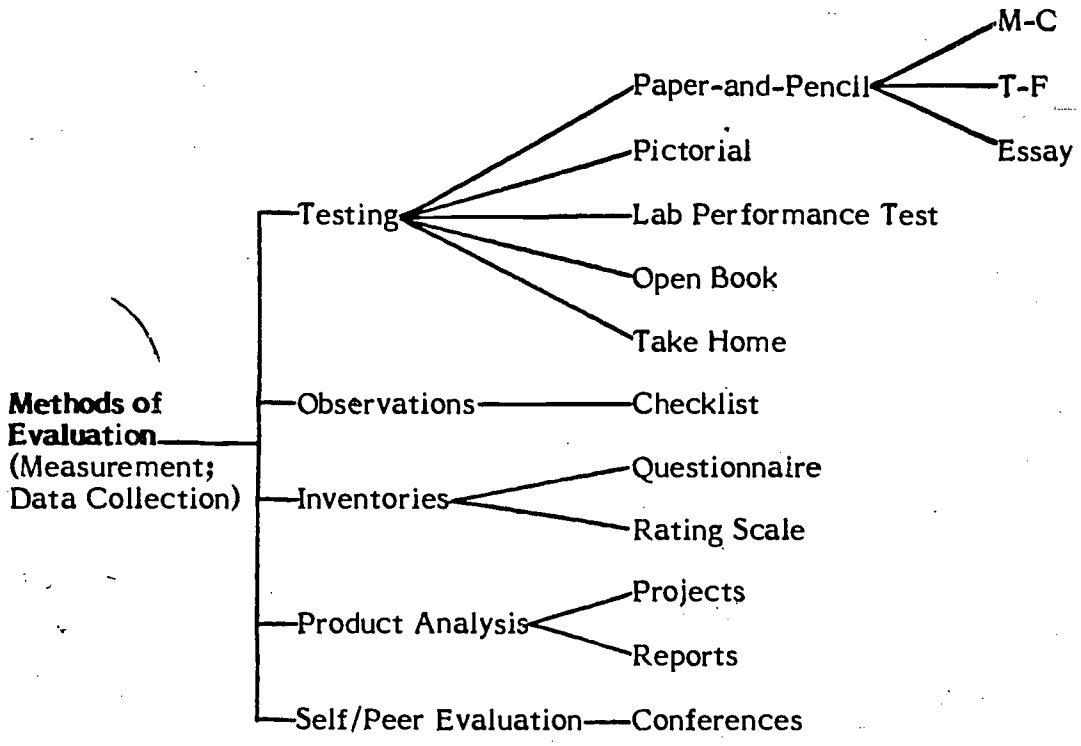
The cognitive domain deals with knowledge and the development of intellectual abilities and skills. The levels of the Cognitive Taxonomy developed by Bloom and associates (13)--Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation--have become part of the common vocabulary of most educators. Data from tests of cognitive

outcomes can be referenced with respect to some comparable group of students (norm) or to a pre-established standard (criterion). These and other ways of assessing outcomes in the cognitive domain are discussed in Chapter 2. In addition to the traditional primary emphasis of education on the cognitive domain, considerable attention has recently been focused on affective objectives. The taxonomy associated with this domain, developed by Krathwohl and associates (53), involves the students' interests, attitudes, feelings, and values. Assessment of these affective outcomes is addressed in Chapter 3. The manipulative or motor-skill abilities of the psychomotor domain--such as titrating solutions and massing objects--are among the various outcomes of science laboratory activities. Measurement of student performance in the laboratory is the focus of Chapter 4.

With the advent of a competency-based teacher certification system and a generous supply of science teachers, the evaluation of teachers is becoming a larger component of the school evaluation program. The most common manner of evaluating teachers has been the analysis of their verbal behavior and of a few selected aspects of their instructional methods. Many other approaches are possible, focusing on such features as strategy of questioning, organization of instruction, and degree and kind of interaction with students.

While many aspects of the classroom (such as socio-emotional environment and interaction) are directly dependent on the teacher, others--such as availability, quantity and quality of equipment, materials, and supplies--suggest a separate evaluational focus on the classroom. Some of the latter items are largely a function of the financial support available through the school district and/or administration. Since the classroom is the site of the actual instruction--where kids and science interact--it is a most important element.

The evaluation of existing curriculum or components of instructional innovation is often undertaken by personnel of funded program assessment projects. This should become a priority item for school district staff. Curriculum evaluation can focus on the ability of a curriculum to accomplish its stated objectives, the efficiency of text or audiovisual components of the curriculum, and the interaction among these variables. Related to these questions are concerns about the logical structure of the instructional materials, their cognitive level, and process orientation.




The measurement or data collection phase is the component most commonly associated with evaluation. Several forms of measurement are listed in the preceding figure, including tests, observations, inventories, product analyses, and self/peer evaluation. As instructional objectives become more varied, measurement devices must become similarly diversified to meet new needs. Some types of data collection are more widely used than others, but all are possible ways of obtaining and recording outcomes either qualitative or quantitative in nature. Of the data collection procedures included in the figure, more examples are offered for the testing category than for any other. Pictorial tests, laboratory performance tests, open book, and take home tests can be used to collect information in addition to the ever-present paper-and-pencil tests.

Podrasky (74) developed tests using 35 mm color slides to present pictorially both the test cases or questions (stimuli) and the responses from which students could choose an answer. Many concepts and principles of science can be illustrated pictorially. This technique serves to reduce the reading demands of tests and can encourage higher level learning. The laboratory performance test is one of the best ways to directly evaluate student ability to make specific observations, measure quantities, work with

experimental apparatus and data, and interpret experimental results. Variations of the paper-and-pencil format, the "open book" and "take home" tests are designed to emphasize homework, independent study skills and the ability to use references.

Each of these modes of assessment has its own characteristics and should be matched according to the demands of the test objective. An obviously inappropriate match, for example, would apply a Likert scale response set (Strongly Agree...Strongly Disagree) to a cognitive question of fact. The major task in making an appropriate and comprehensive match lies in clearly formulating and stating the objective to be tested; once this is done, the best mode of assessment usually surfaces.

In addition to the "testing" mode of data collection, a number of other modes exist. Some objectives (e.g. lab safety) may be best assessed with the help of a checklist or similar device to focus attention on key behaviors or actions. Based on such behavioral evidence, inferences can be made not only about the student, but about the course of instruction. A wealth of information about student performances, preferences, opinions, attitudes, and beliefs may be gleaned from analyses of lab reports, projects, and independent studies. And, although some people doubt the validity of such measures, valuable information can be gathered about students through self and peer evaluation--information that may not emerge in any other phase of evaluation and which is essential to a broadening of perspective.

**Assessment Situations**

The value of a general overview is often blurred without specific instances to bring it into focus. The following hypothetical, yet feasible testing situations are traced through the different dimensions of the measurement and evaluation domain just presented. Besides those dimensions that appear on the outline or in the discussions below, many other aspects of evaluation are involved in concert with the central evaluative thrust. The intention here is not to artificially distinguish between one phase of assessment and another, but rather, by making basic connections among related elements, to suggest further interconnections among elements more indirectly, yet just as significantly, related.

Situation I

Mr. Burke plans to administer an end-of-the-year examination to his biology students in effort to gauge not only how much they have learned about biology during the course, but how well they are prepared for subsequent science courses. The examination will consist of 100 items characterized by the following:

In photosynthesis, the function of chlorophyll is that of:

A. an enzyme in digestion.
B. carbon dioxide in respiration.
C. bile in the digestion of fat.
D. glucose in respiration.

The distribution of all scores will be calculated and plotted, and individual student grades will be determined from the results. Then, both the numerical scores and the corresponding grades emerging from the examination will be recorded on individual grade report forms and entered into each student's permanent file. The information in the files will help the students and their advisors make decisions about what science courses, if any, they might enroll in for the coming year.

The central purpose of the data Burke is collecting is the delineation of the students' grasp of a particular set of material so that the science educator may determine enrollment in subsequent science courses with some assurance that those enrolled can handle the work (articulation). The type of evaluation--a "final"--is summative, and it is administered for the purpose of grading students on a completed unit of instruction. The focus of the evaluation is on the student--specifically, the students' cognitive awareness relative to the subject of biology. Because student grades will be based on the range of scores earned by all students being evaluated, a norm-referenced result will be obtained. The method of evaluation is one of the most common--a paper-and-pencil test comprised of multiple-choice items.

Situation II

On the first day of school, Ms. Sellers opens her General Science course by administering a 50-item scale exemplified by the following:

SELF-CONCEPT IN SCIENCE SCALE

The following statements are to help you describe yourself as you see yourself in science. Please respond to them as if you were describing yourself to yourself. Do not omit any item. Read each statement carefully; then select one of the five responses listed below.

The responses are as follows:

| Completely false | Mostly false | Partly false and partly true | Mostly true | Completely true |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

Remember, respond to the statements as if you were describing yourself to yourself in science.

1. I am satisfied with my ability to make predictions.
2. I do well on number problems in class.
3. I wish I could make better conclusions based on what I have seen in class.
4. I am a person who works well with numbers.
5. I can compare things.
6. I give up when I have to classify things.

She will administer the same scale at the end of the course and will chart the results both for departmental records and for the students to compare. A perceptible change in student self-concept is anticipated.

The purpose of Sellers' scale administration is the monitoring of one dimension of the student's personal growth: self-concept. The initial administration of the scale could be considered of the diagnostic type, a pretest to establish incoming students' attitudes toward science prior to instruction. The focus of the evaluation is, once again, the student, but this time--as the example items clearly show--the assessment centers on the affective components of interests and values. The method of collecting data involves the use of a modified Likert scale, a response format calling upon students to rate a series of statements in terms of degrees of reaction. Inventories of this kind are highly appropriate to eliciting information within the affective domain.

Situation III

Alarmed by reports of plummeting scores on student achievement tests in science, the school board of Technotown--in response to many appeals from its citizens--has commissioned a study of the district's secondary science program in an effort to identify its weaknesses. The study team--comprised of outside consultants as well as teachers and administrators within the system--will use NSTA's Guidelines for Self-Assessment (32) package in their work. The titles of its modules are:

Our School's Science Curriculum
Our School's Science Teachers
Science Student/Teacher Interactions
Science Facilities and Teaching Conditions

Each of these areas will be surveyed by a series of items to which two criteria will be applied: the desirability of the goal and the level of achievement of the goal within the existing program. Each item will be rated on a five point scale for these criteria and their points of convergence or disparity will be plotted by means of a matrix. The results will be compiled and communicated to the school board, who will then present to the school administration the outcomes and indications of the study.

The impetus (or purpose) for the study emerges from the society's concern that the schools produce student-citizens at least minimally literate in matters of science and technology. The type of evaluation employed spans the range from diagnostic to summative, but the principal application of the results will be geared toward remediation of a program found to be inadequate to the goals set for it. The focus of the evaluation is multifaceted, involving the assessment of teachers, classroom facilities, and curriculum components, among many other program elements. The method of evaluation also ranges widely among the modes of observation, inventory, and self-evaluation, with the responses assuming the form of a two-dimensional rating scale, or matrix.

This last test case illustrates just one of many possible ways in which the focus of evaluation may be expanded to assess elements of the instructional process other than those that are strictly student-oriented. Such expansion of assessment objectives constitutes a general trend in science education for which many specific examples may be cited.

## Evaluation and Science Education Trends

Trends specific to science education have been indicated by several science educators and organizations. In Designs for Progress in Science Education, Butts (17) cited the objectives that school science programs must encompass to enhance the survival of our culture. Such related goals as the "understanding of the major conceptual schemes that constitute the basic structure of science," and "the relationship of science to humanities and to social problems which face us now and will persist into the future" will make the task of measurement broader and more complex than it is now.

Looking toward the goal of "Scientific Enlightenment for an Age of Science," Hurd (39) suggested seeking the curriculum which "interprets the scientific enterprise within the broader perspectives of society." He further cited the need for "educating for instability" and the concern with "the development, by students, of sustaining attitudes and values." This proposed vista for science education will require "reordering the subject matter of science, placing it within a cultural context, and demonstrating more concern for human betterment."

Instructional programs revised or created in accordance with these new, more fluid criteria will require similarly modified evaluative techniques which not only take into account new content areas but which also reflect the intent of "opening up" the discipline. Care must be taken, however, that disciplinary standards are not compromised as the scope of considerations expands.

The NSTA position statement on "School Science Education for the 70s" (82) raised several important questions about objectives for the 1970s and their evaluation. This statement recognized the need for supplements to paper-and-pencil type tests, including student self-evaluation; measurement based on criterion performances; evaluation of the higher thought processes; balanced emphasis among different modes of learning and evaluating; and evaluation of objectives in the affective domain. While measurement techniques exist for some of these concerns, the development of additional tools is a challenge for the 1980s.

The following scheme suggests an evolving pattern for measurement and evaluation. The two stages identify a shift from a single level/one-way-street mode of measurement to a combination of modes which is at once multifaceted and multidirectional.

## Predicted Trends in Measurement and Evaluation of Science Instruction

| From..... | To..... |
|---|---|
| 1. Primarily group-administered tests | A variety of administrative formats including large groups, small groups, and individuals. |
| 2. Primarily paper-and-pencil tests | A variety of test formats including pictorial and laboratory performance tests. |
| 3. Primarily end-of-course summative assessment | A variety of pretest, diagnostic and formative types of measurements. |
| 4. Primarily measurement of low-level cognitive outcomes | The inclusion of higher level cognitive outcomes (analysis, evaluation, critical thinking), as well as the measurement of affective (attitudes, interests, and values) and psychomotor outcomes. |
| 5. Primarily norm-referenced achievement testing | The inclusion of more criterion-referenced assessment, mastery testing, and self and peer evaluation. |
| 6. Primarily measurement of facts and principles of science | The inclusion of objectives related to the processes of science, the nature of science, and the interrelationship of science, technology, and society. |
| 7. Primarily measurement of student achievement | The inclusion of measuring the effects of programs, curricula, and teaching techniques. |
| 8. Primarily teacher-made tests | The combined use of teacher-made tests, standardized tests, research instruments, and items from collections assembled by teachers, projects, and other sources. |
| 9. Primarily concern with total test scores | Interest in sub-test performance, item difficulty and discrimination, all aided by mechanical and computerized facilities. |
| 10. Primarily a one-dimensional format of evaluation (e.g., a numerical or letter grade) | A multidimensional system of reporting student progress with respect to such variables as concepts, processes, laboratory procedures, classroom discussion, and problem-solving skills. |

1. Most tests now employed by schools are of the type in which one person, often a teacher or counselor, administers the same test to a large number of students (from one class to several hundred). Although group-administered tests are not likely to disappear, other formats will become more prevalent, such as the testing of individuals or small groups of individuals. As curricula and instructional programs become more individualized, so must the assessment procedures. Johnson (43) and Barker and Frederick (9) have designed computer-based programs which can select items from a "bank" to create many equivalent tests of the same content unit. Teachers may also hand-select items from an available bank of items to meet the needs of individualized testing. For teachers who don't have access to a computer terminal, Farmer and Farrell (23) have suggested a similar technique by which test items are recorded on index cards scored with individualized patterns of holes along the margins for easy identification and retrieval.

2. Although we are at present largely dependent upon paper-and-pencil tests, future assessment procedures will take a variety of forms, including pictorial tests and laboratory performance examinations. The pictorial format can serve to reduce reading demands and to provide a close link to the real phenomena it represents. Students unable to demonstrate their achievements using a paper-and-pencil format may be able to do so if the verbal demands are minimized, whether by working in the laboratory, making a model, or using a mode of response outside the usual range. As teachers, we should be willing to accept any kind of evidence that a student has learned a fact, a principle, or a procedure.

3. The predominant kind of examination students encounter is the end-of-semester or end-of-course summative assessment. The trend is toward the use of additional measures like diagnostic tests, pretests, and formative evaluations. Diagnostic measures can focus on the skills or abilities required to perform successfully in a particular course or unit, such as manual dexterity, spatial perception, and mathematical acumen. Pretests are also administered prior to instruction but focus on variables related specifically to the outcomes of the instruction. Most pretests assess the facts and principles and, less frequently, the science processes included in the instructional materials. If a teacher has detected a weakness in one of these areas, a remedial program specific to the area may be prescribed in

effort to avoid compounding the student's learning difficulties. If this effort is not successful, the student may have to be individually instructed using unique materials or methods. Students who possess, prior to instruction, a high degree of knowledge can be used within the class as teaching assistants, thereby serving to help the teacher and other students as well as themselves. Other alternatives for this kind of student include rapid advance or optional materials study.

Formative evaluation instruments can be used as much for the benefit of teachers and curriculum specialists as for students. Gauging student progress as the instructional unit unfolds can help troubleshoot ineffective teaching techniques or inadequately developed content areas. The encouragement of feedback from the students throughout such evaluation is implicit to the success of its formative aspect.

4. A high proportion of low-level (memory or recall) cognitive outcomes are included in most teacher-made and standardized tests of science achievement. These outcomes are an important part of most courses, but they are not the only objectives. Higher level cognitive test items are harder to construct, requiring much more time and effort to devise. In effort to help overcome some of these difficulties, Chapter 2 offers several suggestions. There is no magic formula to determine the appropriate distribution of various levels of objectives; this distribution will vary with the nature of the course: its goals, students, and teachers. Science courses which aspire to such goals as thinking critically, interpreting data, and formulating hypotheses should include tests that measure outcomes above the recall level.

The goals of many science courses include statements about student interest in and appreciation of science and scientists and, increasingly, concerns about the relationship of values to science and technology. These affective outcomes must be assessed and monitored, albeit in different ways than are cognitive outcomes. Measurement of affective objectives is discussed in Chapter 3.

One unique aspect of science instruction is the entire system of experimental inquiry, involving an emphasis on laboratory procedures and skills, and a reliance on data and replicable evidence. Relatively few science courses, however, include attempts to assess student ability or achievement in the laboratory. Of this domain, the least measured part is

the psychomotor or manipulative portion. Chapter 4 presents suggestions and samples for assessing outcomes related to science laboratories.

5. The frame of reference for the vast majority of past and present assessment procedures is a "norming" group. With standardized tests, this norm group might be a sample randomly selected from a national population. With tests developed by classroom teachers, commonly adopted norm groups include a single class; a group of classes under a single teacher's direction; and a group of all classes in the school or district. The performances of individual students are compared to the performance of their peers by means of some kind of norming group. A trend in evaluation is toward the specification of objectives for a particular unit of study, including the level or standard of performance to be achieved for each objective. These objectives and their criteria fit ideally into a measurement system through which teachers could describe expected student outcomes. This new frame of reference for evaluating student performance has been called criterion-referenced measurement. Criterion-referenced systems and their applications to student grading are discussed in more detail in Chapter 6.

6. The "facts and principles of science" comprise a major portion of the goals and outcomes of school science programs. This is partially due to the limitations imposed by the explosion of knowledge within each of the science disciplines, as well as the perceived dependence of college science and engineering courses on this core of knowledge and understanding from high school science programs. A wider spectrum of topics is being recommended as more appropriate for the majority of high school students, more of whom will become scientifically literate citizens than will become scientists or other science professionals. Included in these recommendations are concerns about the processes of science, the nature of science, and the interrelationship of science, technology, and society. Both the cognitive and affective domains are applicable to these concerns.

7. The focus of measurement and evaluation activity has historically been student achievement, but interest in the evaluation of curricula, programs, and teaching techniques is growing. The ultimate goal of such multifaceted evaluation will be the further assessment of how each element relates to and influences the evaluation of the other. This will serve to distribute the responsibility for the educational process more evenly among the teaching and learning factions than it has been in the past.

8. At present, teacher-made tests outnumber--in terms of those developed as well as those actually used--tests from all other sources. With the upsurge of interest in the process of evaluation, however, many individuals and groups have become involved in developing and informally disseminating evaluation instruments that may be effectively used in more than one setting or context. Summaries and reviews of these instruments often appear in educational journals or newsletters, and Mayer (64) has compiled an excellent source book on unpublished evaluation instruments in science education.

At the same time, as computer and microcomputer facilities become more accessible, standardized tests and their results will be integrated into the day-to-day evaluation activities of teachers interested in a variety of perspectives on their students' progress.

9. The results of tests are usually summarized and reported in terms of total test scores. Attention is beginning to focus on such additional data as sub-test scores, individual item scores, and indices of difficulty and discrimination for each item. These data are obtainable through hand calculations, and are part of the information provided by most computerized item and test scoring programs. Most of these programs work with either optically scanned answer sheets or computer cards, and results are rapidly tabulated by most computer facilities. As such facilities become more widely available and understood, a vast array of information will be provided to enable teachers to supplement their evaluative criteria beyond the total test score. These procedures are discussed in more detail in Chapter 5.

10. Students are commonly evaluated by a single grade or number which is a reflection of their total performance in the science classroom. Multidimensional systems are being developed for describing and communicating student progress in the areas of concepts, processes, laboratory procedures, problem-solving skills, classroom interactions, and various affective variables. As the goals and outcomes of school science programs become more complex, the dimensions of the measurement task will similarly increase in complexity. Examples of forms that assessment might assume include the use of written evaluations in addition to quantitative reports; the involvement of the student in self-evaluation exercises; and the engagement in conferences of students, teachers, parents, and peers in effort to form a more total picture of progress and performance.

Although these ten trends or predictions may appear to be independent, many forms of interaction exist among the categories, with development in one area enhancing or hindering progress in another. This kind of interaction yields a variable rate of development among the categories and thus further increases the complexity of the task ahead of us.

**Implications for Science Educators**

The institution of these trends will exert additional pressure on the accompanying management system. The teacher's grade book is, at present, the repository for most of the information on which student grades, advancement, and achievement are based. Computer facilities could easily accomplish the required storage and retrieval of this information, but the apprehensions and misconceptions of students, parents, and the community will have to be addressed before such a system can be successfully implemented. In addition, specific safeguards will have to be provided to prevent misuse and abuse of the system (e.g., invasion of privacy).

Science educators at the local level must develop a variety of instruments with which to survey the status of their existing science programs as well as to convey their findings to other school personnel, students, parents, and the community. Local needs and resources must be identified by and communicated to those responsible for them and those most responsive to them. Especially considering the ever-increasing demand for accountability by the "back to basics" contingent and others, science educators must prepare a well-documented rationale and defense not only for new programs, but sometimes just to maintain existing ones.

We science educators know best what the goals of our programs are and how to assess their outcomes. If we don't devote the time and effort to make our assessment tools valid and specific to our goals, our programs will be evaluated by someone else—on their terms and according to their priorities. In the science laboratory, a well-developed question is often the most important factor in finding a solution or, at least, in reaching a resolution. The same is true for the science classroom and curriculum.

Some of the basic parameters of measurement and evaluation in science education have been identified here and some probable trends have been indicated. As these trends continue and mature, the community of science educators is urged to equip itself for effective participation.

# CHAPTER TWO

# Assessing Cognitive Outcomes in Science

## Introduction

Item:
What do teachers complain about almost as actively as do the students?

A. The weather.
B. The principal.
C. The classroom test.

Answer C--though not the only answer--is at least the option over which both teachers and students have most control. Just as teachers grumble about preparing classroom tests, students grumble about taking them, and both send up quite a howl about the grading process.

Few teachers have had substantial instruction in constructing tests, although tests are perceived to be a highly critical part of a course by teachers and students alike. In the absence of clearly defined course objectives, the test may emerge as the statement of what is "really important" in the course, a statement which is undercut by the inclusion of trivial or picayune details as much as it is by over-generalized "giveaways" or whimsical emphases on the teacher's pet topics. And sometimes, the test may be employed as a punitive device ("I'll throw something in that none of them can answer") or as a snare ("They were supposed to have read this material even though we never mentioned it in class"). The test, unfortunately, is different things to different people, so it isn't difficult to

understand why numerous complaints are generated about the testing process at all instructional levels and from each instructional standpoint.

Some of these differences can be resolved by the careful application of a few widely acknowledged principles and techniques of successful test making and administration. These principles, however, are not panaceas. Writing good test items is a creative, artistic endeavor that requires (a) an excellent command of the content to be tested, (b) a comprehensive grasp of the behaviors to be evaluated, (c) a thorough understanding of the students' backgrounds, abilities and interests, and (d) a precise understanding of the English language.

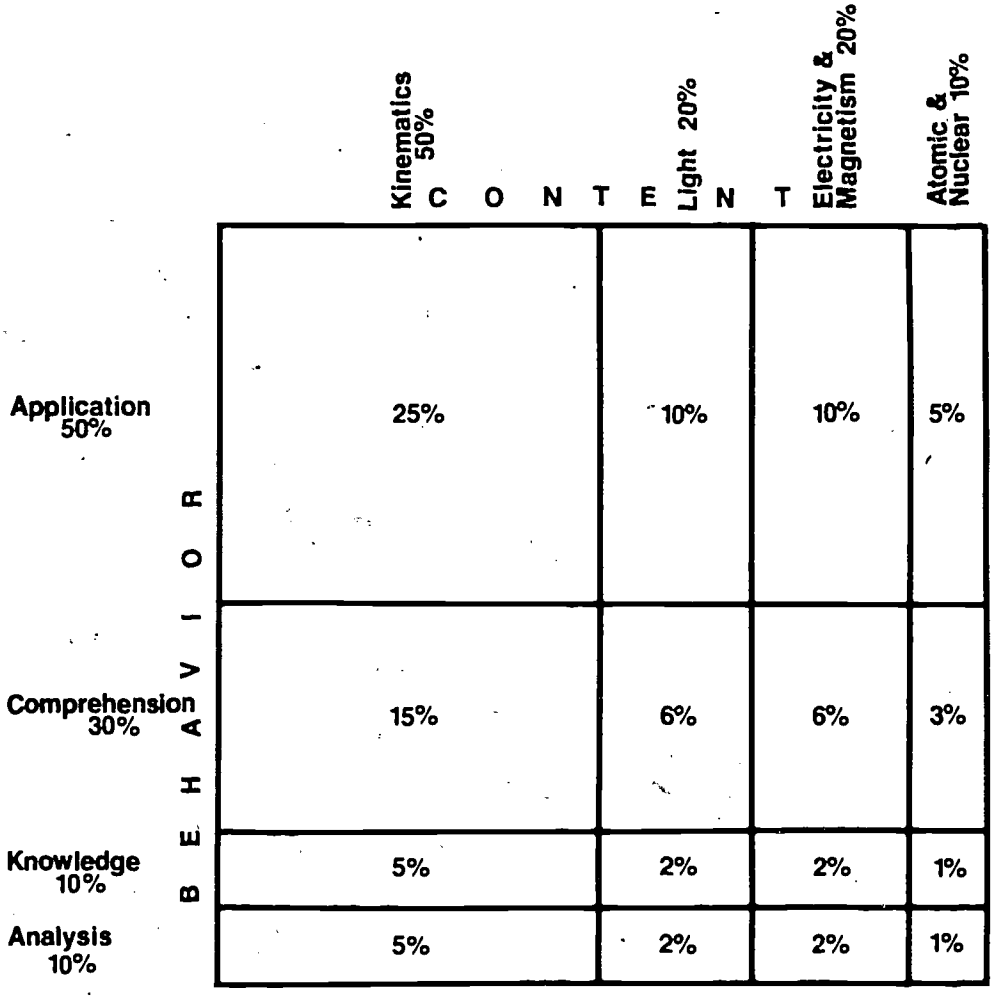
**Delineating Objectives**

Planning and designing a test require a clear understanding of the objectives to be tested. For this purpose, a precise statement of expected student outcomes from a particular unit or course will be most helpful, especially for the novice test-maker.

Objectives are written at a variety of levels with differing degrees of generality or specificity, depending on whether they are for an entire school science program or one daily lesson. Goals for a science program or an entire course are necessarily general and, therefore, are not likely to be written in behavioral terms. Objectives for individual lessons can easily be stated in behavioral terms and thereby contribute meaningfully to both instruction and evaluation. Regardless how teachers construct items for evaluating a given unit of instruction, they are implicitly or explicitly conceptualizing what students should be able to do after the unit of instruction has been completed. According to Mager (61), objectives should be identified which (a) are properly stated in terms of student behavior, (b) include conditions under which the behavior will be expected to occur, and (c) state the performance standard (criterion) of student behavior. Anderson (5) contends that most educational researchers have yet to meet the "primitive first requirement, namely that there is a clear and consistent definition of the things being counted." Anderson's requirement is at least partially satisfied by a table of specifications.

A very helpful method for organizing the objectives of a unit or course is a table of specifications (TOS), usually a two-dimensional chart including dimensions of content and behavior. A TOS is especially helpful for

developing a balanced, fair and relevant examination. The example in the figure below illustrates how the two dimensions and the proportional emphasis among the categories can help in constructing or selecting specific items for a test on physics topics.




In this hypothetical example, the application behavior is deemed to be the most important student outcome--representing 50% of the expected student behaviors. Comprehension skills are next most valued with 30% of the total behavior dimension, and behaviors of knowledge and analysis are each weighted with 10% of the total. In this example, the most important content area is "Kinematics," which accounts for 50% of the content on the table. Each of the topics "Light" and "Electricity and Magnetism" is to be assessed by 20% of the items, and the least-stressed category is "Atomic and Nuclear," with a 10% emphasis.

The content dimension can be apportioned into varying percentages of total assessment by examining the class time spent on each category or the relative emphasis in a course outline or curriculum guide. The behavior dimension is not nearly as easy to apportion, but must be based on the teacher's subjective judgment in conjunction with goals and objectives that are part of the course outline or curriculum. Several revisions may be necessary before a realistic allotment of behaviors for the exam of a particular unit or course is established.

By cross multiplying the column and row proportions, the percentage of items for each "behavior-content" objective can be obtained. For example, "Comprehension-Light" objectives comprise 6% (.30 x .20 or .06) of the items on the entire test. Similar computations produced the values of each of the "boxes" in this sample table of specifications.

**Describing and Organizing Behavioral Outcomes**

The most frequently used scheme for describing cognitive behaviors is the Cognitive Taxonomy developed by Bloom, et al. (13) with its levels of knowledge, comprehension, application, analysis, synthesis, and evaluation. These levels were selected as appropriate for general educational objectives, not just for science. Bloom's scheme has been widely used in educational research as a tool for constructing and analyzing exams and as an aid for curricular and instructional materials development. Other schemes for organizing science teaching objectives have focused on processes and problem solving skills.

The BSCS Test Grid Category system was designed by Klinckmann (48) as an aid for constructing BSCS exams. The first category, "Ability to Recall and Reorganize Materials Learned" is, for the most part, identical to Bloom's knowledge category. The other BSCS categories were constructed to fit the kinds of behaviors unique to science classes, and therefore they are of special interest to science teachers. An overview of these categories, taken from the BSCS Newsletter, is included opposite for reference.

Several methods have been discussed for possible use in organizing behavioral outcomes of science courses. If a specific objective with behavior and content components is examined, a particular item format (e.g., essay, multiple-choice) may emerge as the one most appropriate to the considerations being made. The thinking processes involved in fulfilling an

PAGES 23-24 "BSCS TEST GRID CATEGORIES" REMOVED
DUE TO COPYRIGHT RESTRICTIONS

objective should determine the selection of an item format requiring the student to utilize a similar process in answering the item correctly. For instance, if the objective involves the ability to make choices among various courses of action, the multiple-choice format may be most appropriate. On the other hand, if the objective stresses the ability to make an original analysis of an issue or to synthesize several positions, the essay format may be the best option. For some objectives, several item formats may seem equally suitable. Then the teacher may wish to write items using several different formats and evaluate the effectiveness of each in terms of how well it serves to elicit a valid student response.

### Creating a Test Item Pool

Some standardized achievement tests include worthy items, but individually constructed items are recommended for several reasons. Such items may be structured to closely parallel the specific objectives of a particular course, and the experience gained from constructing the items can facilitate the later revision and selection processes. The development of good tests is hard work--so hard that many people don't want to discard good items after using them only once. Preparing a new set of items for each test is not only taxing, but also inadvisable from the standpoint of maintaining quality control over test production. According to Sawin (81),

> . . . even the professional test developers employed by test publishing companies cannot produce high quality items rapidly. In an eight hour day, such writers may turn out approximately twenty items measuring factual information. They may, however, spend all afternoon on a single item that measures a higher level of ability.

One solution to this problem is the creation of a test item pool (sometimes called a file or bank) from which items of different kinds may be chosen to assess particular objectives.

Items should be coded for behavior and content components so that items can be selected which produce a balanced, representative test. Ideally, the larger the pool the better, but certainly a pool size of three or four times the number of items on any particular test is a good beginning. As student responses yield test data, items may be added, deleted, or revised. For efficient storage and access, each item may be written on an index card, allowing room for coded information about the item. Such

information might include the behavior and content components addressed by the item; the correct answer; the source of the item; and comments emerging from the item's use.

Although the establishment of an item pool is a lot of work, it can prove to be most worthwhile. If several teachers are teaching the same course, a group effort in developing the item pool could be helpful and challenging. Teachers might contribute their best items and, in turn, would have access to the items contributed by other members of the group.

### Writing Essay Items for Science Tests

Essay testing, an outgrowth of individual oral testing, was originally justified as being more impartial and reliable than oral testing, and certainly more efficient with classes of more than ten students. The essay item (sometimes also called short answer, open-ended or problem-solving) is the major type of "supply" item for which students must provide the answer, rather than recognize or select the correct answer from several choices provided. Another characteristic of the essay format--the relatively subjective way in which it must be scored--has prompted comparisons with item formats which are more objectively scored by means of some predetermined system. A chart from the ETS booklet, Making the Classroom Test (62), summarizes these comparisons and appears opposite.

There are many types of essay questions eliciting a variety of behaviors. Essay questions can be used to assess recall, understanding, and judgment behaviors, and they are ideally suited to testing higher level objectives like the organization and synthesis of knowledge. Many guidelines to aid in the development of good essay questions have been produced. The following discussion is based largely on ideas developed by Marshall and Hales (63) in their handbook, Classroom Test Construction.

(1) Allow adequate time for the construction of items. Although essay items are somewhat easier to write than are some other types of items, they still must be carefully constructed to be useful. Before choosing the wording of the question, consider carefully the content and behavior to be tested as well as the backgrounds of the students. Allow time for several revisions.

(2) The problem should be defined explicitly. In the course of writing an essay item, expectations of the nature of the answer often emerge in the

| ESSAY | OBJECTIVE |
|---|---|
| **Abilities Measured** | |
| Requires the student to express himself in his own words, using information from his own background and knowledge. | Requires the student to select correct answers from given options, or to supply answers limited to one word or phrase. |
| Can tap high levels of reasoning such as required in inference, organization of ideas, comparison and contrast. | Can also tap high levels of reasoning such as required in inference, organization of ideas, comparison and contrast. |
| Does not measure purely factual information efficiently. | Measures knowledge of facts efficiently. |
| **Scope** | |
| Covers only a limited field of knowledge in any one test. Essay questions take so long to answer that relatively few can be answered in a given period of time. Also, the student who is especially fluent can often avoid discussing points of which he is unsure. | Covers a broad field of knowledge in one test. Since objective questions may be answered quickly, one test may contain many questions. A broad coverage helps provide reliable measurement. |
| **Incentive to Pupils** | |
| Encourages pupils to learn how to organize their own ideas and express them effectively. | Encourages pupils to build up a broad background of knowledge and abilities. |
| **Ease of Preparation** | |
| Requires writing only a few questions for a test. Tasks must be clearly defined, general enough to offer some leeway, specific enough to set limits. | Requires writing many questions for a test. Wording must avoid ambiguities and "giveaways." Distractors should embody most likely misconceptions. |
| **Scoring** | |
| Usually very time-consuming to score. | Can be scored quickly. |
| Permits teachers to comment directly on the reasoning processes of individual pupils. However, an answer may be scored differently by different teachers or by the same teacher at different times. | Answer generally scored only right or wrong, but scoring is very accurate and consistent. |

writer's mind. But too frequently, a student has to be a "mind reader" to figure out the problem to which he is to respond. An essay item is not valid if students do not interpret the question the same way. With varying interpretations, students are responding to different questions, making evaluation difficult at best. A colleague could provide valuable insight by critiquing each question with an eye to its possible interpretations and by eliminating ambiguity and awkwardness in the wording of the item.

(3) The problem should be limited. It is very difficult for a student to respond adequately to a question covering a large content area, and often only broad, unsupported generalizations are elicited by such a question. An unlimited question--like "Discuss Photosynthesis"--invites random "cranking out all you know" and outright guessing, both of which lower the validity of the item. Students should be guided on the level and focus of such discussion questions, e.g., "Discuss the dependence of the chemical processes of photosynthesis on environmental factors."

(4) The directions for essay items should be stated explicitly. Students must know precisely what is expected of them; how much time to spend on each question; the type of information to be included in the responses; and the form in which the responses are to be written. A statement describing the relative weight of each question should be included either in the general test directions or as a part of each question. Exemplary directions follow:

DIRECTIONS

> Please answer each of the following five questions. Answers must include explanations which describe the cellular mechanism involved. Each answer should be less than 250 words (one page). Each is worth ten points. Two of the ten points will be used to evaluate the communication skills used in the answer, i.e., sentence structure, punctuation, and spelling.

(5) Do not ask optional questions. According to Marshall and Hales (63), "every question used in an examination should be important and therefore should be answered by every student." If different students have responded to a different set of items, they have, in effect, taken a different test. 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.

(6) The conscientious scoring of essay items is among the most time-consuming and frustrating tasks of teachers. Construction of a detailed key for scoring responses to each question is a necessary first step in this area. After writing a question, the teacher should write what is considered to be a model answer to the question. In addition to improving the scoring reliability of the exam, this process will help the teacher spot ambiguities or inconsistencies in the item. A model answer becomes the criterion against which each student's response will be judged. Without such a criterion, results may be affected by the teacher's unconscious, subjective response to such extraneous factors as the "halo effect," handwriting, or verbosity. A simple way to minimize one form of subjectivity is to ask the students to write their names only on the back of the test papers, allowing the teacher to score them anonymously. A second recommendation is to score each item for all students at one sitting, instead of evaluating the entire test for each student at one time.

The instructor may "partition" the model answer into a series of points or features, each of which is specifically described. Each element in the answer is then assigned a number of points, depending on the instructor's judgment of its centrality to the total answer. If used consistently, this method can yield consistent, reliable scores.

Since it is difficult to grade essays reliably, some teachers are more concerned with offering comments than assigning number or letter grades. Specific written comments can be as helpful as a grade in communicating to students their strengths and weaknesses.

## Writing Completion Items for Science Tests

Completion items represent a compromise between essay items and objective items. In completion items, the students must complete a statement by writing the answer(s) in the space(s) provided. For example:

The formula for Methane is_____.

The student is required to supply the answer rather than select it, so the emphasis is primarily on recall. It is quite difficult to write completion

items measuring higher levels of cognitive ability. "Guessing" is of little consequence in completion items. If the items are well-constructed, the answers will be well-defined and can be scored rapidly and reliably. Since a large number of these items can be completed per classroom period, it is possible to assess student knowledge of a broad spectrum of content.

A major difficulty with completion items is constructing them so that only a single answer (or small set of answers) is considered correct. Poorly written completion items are those for which a diversity of responses could be considered correct. The wording of such items must be improved or the scoring becomes very difficult and the scores less reliable.

Information based on principles from Improving the ClassroomTest (40) is presented here as a guide for constructing completion items.

(1) Avoid vague items that don't clearly limit the answer to one or two specific words or phrases. Minimize the opportunities for students to misinterpret the question. It is very difficult to determine if such misinterpretations are honest mistakes or if the student is "putting you on." For instance, consider the following item:

Matter occurs in the 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 _____.

(2) Do not require more than one or two completions in any one item. The following item is an example of such a multi-mutilated item:

Most____green____plants produce____sugar____
from____$H_2O$____and____$CO_2$____.

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

(3) Place the blank at or near the end of the statement. When the blank is at the beginning of a sentence, the student must read the statement and then retrace steps to decide what should be written in the blank.

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.

(4) 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 by either using the "a/an" phrase or 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,"

(5) In general, the use of the completion form should be avoided when other forms are more appropriate to the objectives and learning processes being tested. The basic purpose of completion items is to determine if students can recall a particular word or phrase, in contrast to having them recognize it among a group of distractors. Some items, however, assume the form of a completion item without incorporating its intent. The following

item is an example of an inappropriate use of the completion form:

The density of a floating body is_____than 1.

This item is basically a true-false item or two-option multiple-choice item, because only two answers are viable--either "greater" or "less." For an item of this kind, the completion format offers no special advantages to recommend its use--recall is not the operative objective or ability being assessed, and the guessing factor is not limited, but rather increased.

Several revisions of this item are possible, each of which would yield a more efficient use of testing time while at the same time increasing the degree of objectivity in scoring the item:

a. Make it a two-option M-C item with options "greater" and "less."

b. Make it a three-option M-C item with options "greater than," "less than" and "equal to."

c. Make it a True-False item with either "greater" or "less" in the basic statement.

d. Make it a better completion item by eliminating the words "than 1" from the original statement.

(6) In computation problems, specify the degree of precision expected. Often when the emphasis is upon method and comprehension, problems are written so that the answers come out easily and evenly. If the answers involve fractions, decimals or approximations, students should be told what degree of accuracy will be expected when answers are scored. For instance, 3; 3 1/7; 3.14; or 3.1416 could all be considered correct in answer to this completion problem:

If the radius of a circle is one inch, its area is_____.

Similarly, students should be told if the unit of measurement must be included in the answer to be considered correct. In the above item, "square inches" must accompany the number to be considered correct. If in a particular item the computation is the main concern, the units may be included in the statement itself:

If the radius of a circle is one inch, its area is_____sq. in.

**Writing Matching Items for Science Tests**

A matching item consists of a list of stimuli and a list of responses. The student must select the response that is most closely related to each of the stimuli. A sample matching item follows:

DIRECTIONS In the space next to each chemical formula in Column A, write the number from Column B that represents the compound indicated by each formula.

| Column A | | Column B |
|---|---|---|
| _____a. | $H_2O$ | 1. Hydrogen |
| _____b. | $CO_2$ | 2. Water |
| _____c. | CO | 3. Carbon Dioxide |
| _____d. | $H^2$ | 4. Methane |
| _____e. | $CH_4$ | 5. Hydrogen Peroxide |
| | | 6. Carbon Monoxide |
| | | 7. Octane |

Matching items require little reading time, so many questions covering a broad range of content can be used in a class period. Scoring these items is simple and direct and the guessing factor is minimal. They are effective for assessing student knowledge of facts, principles and relationships between one set of objects and another. Matching items are very efficient for measuring the connection between names, dates, categories, classifications, symbols, equations, and formulas, as well as sequences, methods, and processes. They are not well suited for assessing the higher level behaviors such as analysis and interpretation. The following suggestions will aid in constructing matching items:

(1) Within each item, be sure the stimuli and responses are homogeneous. When the stimuli and responses are heterogeneous, the item measures only superficial verbal association and can be solved with limited understanding. In the following example from Improving the Classroom Test, each item in Column A is so obviously related to one of those in Column B that the others become totally implausible distractors:

| Column A | | Column B |
|---|---|---|
| _____a. | lever | 1. block and tackle |
| _____b. | gas | 2. carbon dioxide |
| _____c. | pulley | 3. crowbar |
| _____d. | solid | 4. brick |
| _____e. | kinetic energy | 5. $9.8\ m/sec^2$ |
| _____f. | acceleration of gravity | 6. moving car |
| | | 7. Newton |
| _____g. | planet | 8. Mars |

(2) Keep the lists of stimuli and responses relatively short. Lists with 25 items become very time-consuming if students must make 25 x 25 comparisons. The optimum size is between five and ten items. It is difficult for a teacher to maintain homogeneity in a long list of items (except in very trivial examples, like matching names or symbols for the chemical elements).

(3) Arrange the lists of stimuli and responses for maximum convenience and clarity to the students. Most students first read the stimuli in the left column and then scan the responses in the right column to find the match. Assuming this procedure, students read the stimuli only once while the responses will be scanned several times, so the longer, more complex statements should be placed in the left column. Shorter and simpler statements should be placed in the right column since they will likely be read several times. Additionally, it is helpful if the individual stimuli and responses are arranged in some logical order to simplify student scanning and searching. For example, numbers and dates can be arranged in chronological order while names and most verbal responses can be alphabetized. The simpler and clearer the tasks become for the student, the more useful the item is likely to be.

(4) Explain clearly in the directions the basis upon which the items are to be matched and the procedure to be used. In most matching exercises the basis for the matching is obvious, but a general policy of always stating precise, explicit directions for each item is recommended. In the illustrative "poor" item below (also from Improving the Classroom Test), students could become quite confused about the basis for matching stimuli and responses, and some students might not be able to determine the intended basis for classification:

POOR On the line at the left of each item in Column A write the number of the matching item in Column B.

| Column A | | Column B | |
|---|---|---|---|
| _____ | a. lens | 1. | barometer |
| _____ | b. mercury | 2. | electric light bulb |
| _____ | c. vacuum tube | 3. | gasoline engine |
| _____ | d. electromagnet | 4. | microscope |
| _____ | e. filament | 5. | periscope |
| | | 6. | radio |
| | | 7. | telephone |

BETTER The following columns refer to several "minerals" and their "uses." You are to match each mineral with its primary use. Place the letter of the use in the blank space preceding each mineral.

| Column A - MINERALS | | Column B - USES |
|---|---|---|
| _____1. | diamond | a. production of aluminum |
| _____2. | slate | b. medicine |
| _____3. | sandstone | c. photoelectric equipment |
| _____4. | hematite | d. jewelry |
| _____5. | bauxite | e. production of iron |
| | | f. roof shingles |
| | | g. production of copper |
| | | h. buildings |

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 single stimulus. Specific directions and sample items should be used if this is the student's first exposure to this kind of matching procedure.

(5) Provide extra responses to reduce the effect of guessing. If the same number of stimuli and responses is used, students may be able to make some correct matches by a process of elimination. To minimize the possibility of giving "free matches," at least two or three extra responses should be offered. To be effective distractors, these items must be homogeneous to the other responses. Another alternative is to allow some responses to be used more than once and others not at all. Again, this kind of stipulation should be clearly set forth in the directions to the student.

(6) The entire matching exercise should appear on a single page. When part of an exercise is on one page and part on another, students are unduly confused by the flipping back and forth required to respond to the item. If necessary, leave part of a page blank rather than split up a matching exercise.

**Writing True-False Items for Science Tests**

True-False (T-F) items are widely used by many teachers. Arguments supporting this widespread use focus on the easy construction, objective scoring and broad sampling possible with T-F items.

As with all items, some disadvantages have been noted. A major limitation of T-F items is that they measure primarily factual information

and are not very well suited to higher level objectives. T-F items are too often ambiguous and thus measure reading ability rather than the intended content. Statements that are absolutely true (or false) with no qualifications or exceptions are hard to construct. Often it is the better-informed students who are confused by such items, since they see the need for qualification and question the absoluteness of the truth. For instance:

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 seawater. While most true-false items appear to be very simple, they are often open to many interpretations and questions.

Another major limitation of T-F items involves the large guessing factor. Without even reading an item, a student can flip a coin and have a 50% probability of choosing the correct answer. The guessing factor can be reduced by increasing the number of items in a test, as shown by this chart from Improving the Classroom Test:

| Number of True-False Items | Chances of Answering at Least 70% Correctly by Chance Alone |
|---|---|
| 10 | 1 out of 6 |
| 25 | 1 out of 50 |
| 50 | 1 out of 350 |
| 100 | 1 out of 10,000 |
| 200 | less than 1 out of 1,000,000 |

The following guidelines, adapted from Improving the Classroom Test, are offered for writing T-F items:

(1) If the answers are to be indicated on the test paper, it is recommended that students circle their responses, usually T or F. This is superior to having students write out T or F, t or f, + or -, + or O, or true or false. With hasty and sloppily written responses, the instructor may find it difficult to distinguish between such marks, especially if erasures or cross-outs are permitted. Of the above pairs of symbols, the use of + and O is best, but it is still not as definitive as having students circle one of the responses printed on the test page. If printed answer sheets are used, the general convention is to have the student mark A for true and B for false, if it is not possible for T and F to be printed on the sheets. Students must be informed of the response system to be used.

(2) The T-F item should be stated clearly. If a T-F item is open to interpretation, in a class of 30 students, at least one will interpret the statement in a way other than that intended. The student then faces the problem of guessing which meaning is intended. An incorrect guess may result in credit being denied a student who, in fact, knows the answer to the question intended by the teacher. For example:

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 an answer contrary to the one that is expected.

Items should be stated clearly and specifically but should not be word-for-word excerpts from the textbook. When taken out of context, such statements can be ambiguous. Such a practice also reinforces questionable study habits and projects the text to be the ultimate authority rather than a source for facts and a tool to help learn how to use and extend ideas.

Qualitative terms such as "large," "many" and "better" should be avoided because they are relative and can be interpreted differently by different people. Whenever possible, specific quantitative language, such as "80%" and "14 pounds," should be used.

(3) Highlight the central point of an item by placing it in a prominent position in the statement. Students should be able to locate the crucial aspect of the question rapidly. If the key words are hidden from the students, reading ability, IQ, or "test-wiseness" is being tested instead of the intended outcomes. One way to avoid this problem is to eliminate double-barreled or multi-barreled items that are often 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 That evaporation is a cooling process explains why 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 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. It is

generally recommended that such compound items be split into several items, separately testing each idea.

(4) Avoid items that can deceive students. A common but questionable practice is the use of trick questions, like the following example from Improving the Classroom Test:

T F The chief component in the automobile engine combustion is nitrogen.

The answer was intended to be "true" since nitrogen constitutes almost 80% of the air drawn into the chamber. Students who realize that nitrogen itself does not enter into the combustion process (only gasoline and oxygen do), however, may get the answer wrong. If the issue is whether students know which gas is the major component of air, the following question is a direct way to find out, and is certainly more ethical:

T F Air is composed primarily of nitrogen.

Trick questions can negatively influence student attitudes toward tests, science and perhaps school in general.

Another kind of deceptive item is one with lots of "window dressing"--nonessential information added to an item for one of a variety of reasons. Sometimes the purpose is commendable (such as to make the item more interesting or relevant to the pupil) but that purpose is defeated if the measurement objective is obscured. Great amounts of window dressing can begin to put a burden on the student who is deficient in reading speed or comprehension. Statements which include "double negatives" can be extremely deceptive and confusing to students. Such items stress intelligence or test-wiseness, instead of the content to be tested. There are instances when negative statements are desirable and appropriate, especially if the concept tested is essentially a negative one:

T F Iodine should not be applied around the eyes.

In such cases, it is recommended to alert the student to this by capitalizing and/or underlining this key word so as not to be intentionally deceptive:

T F Iodine should NOT be applied around the eyes.

(5) Avoid words which give irrelevant clues to the answer. These words, often called specific determiners, enable the student to answer correctly without possessing the specific knowledge intended to be measured. Two examples follow:

POOR T F The brightest stars are always the closest.

POOR T F No picture--no sound in a TV set may indicate a bad 5U4G tube.

Both items are poor because of the specific determiners "always" and "may." Students with little knowledge of astronomy or electronics would still have a good chance of answering the items correctly. The best guess with most items using "always" is that they are false, and the best guess with items using "may" is that they are true. Studies have shown that statements incorporating strong words such as always, no, never, all and none are most often false, while statements with moderate words such as some, may and often are generally allowed to be true. If "clue" words can't be eliminated altogether, care must be taken to balance the number of true and false items using each kind of "specific determiner."

(6) Construct an approximately equal number of items keyed "true" as keyed "false." Proportions of true and false items can vary from test to test, but consistent trends will be detected by students and may be used to help make decisions on some items, thereby detracting from the validity of the intended measurement. Care should also be taken to sequence items so that no regular pattern of "T" or "F" occurs within the test. Attending to these details can aid in making the measurement as valid as possible.

In addition to the regular T-F items, two adapted forms have been suggested for potential use: (1) Modified True-False, and (2) Multiple or Cluster True-False items.

In Modified T-F 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 modification:

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~~.

A modification in scoring is also possible with this option; one point could be allowed for correctly identifying the item as true or false and then another extra point for supplying the correct answer for false items. If this scoring method is used, students must be thoroughly informed about it. These kinds of items do require more time to complete, so fewer items can be scheduled per unit time.

A similar modification requires students to state the reason why the statement is true or false. Scoring then could be expanded to a four point scale encompassing the combinations of right and wrong answers and right and wrong reasons. Citing why some true statements are true without falling into a pattern of circular reasoning, however, sometimes presents undue difficulties for students.

The last example of modified T-F items involves using a three point scale with a middle "hedging" category. Phrases that have been used in this kind of modification include:

1. True Uncertain False
2. Correct Partially correct Incorrect
3. Agree Undecided Disagree
4. Yes It depends No

The validity of this kind of item may be increased by allowing students to additionally communicate "what it depends on" or "what makes it partially correct," etc. Otherwise, the middle response may serve simply to multiply the "guessing" factor.

Multiple or Cluster T-F items look like multiple-choice items except that students can select none, one, several, or all the responses as "correct." A cluster T-F item format may be a suitable substitute for a multiple-choice item for which homogeneous distractors are difficult to develop. The following is an example of a cluster T-F item:

The current through an appliance can be increased by:

T F A. Increasing the voltage across the appliance.
T F B. Decreasing the resistance of the appliance.
T F C. Adding another appliance in a parallel circuit with the appliance.
T F D. Adding another appliance in a series circuit with the appliance.

**Writing Multiple-Choice Items for Science Tests**

Almost every American citizen who has been educated beyond the third grade has experienced the multiple-choice (M-C) item. It is widely

used, especially in standardized achievement tests. M-C items can test a wide range of behaviors from recall to the higher level skills. Unless the items are very complex or involve large reading passages, many items can be completed per unit time, thereby providing a good sampling of content in a single test. M-C items can be scored rapidly and objectively and have a much smaller guessing factor than T-F items. The chance of "guessing" correctly on individual items is based on the number of alternatives: 50% with 2 choices, 33% with 3 choices, 25% with 4 choices and 20% with 5 choices. These percentages are based on the assumption that all responses are plausible.

One complaint about M-C items is that they are restrictive and stifling to creative students. This may be true of poorly written M-C items, as with poorly written items of all formats. Well-written M-C items, however, can be challenging and fair to most students. It is good to keep in mind that an entire assessment system should not be based on any one kind of item, no matter how good that format may be. A variety of item formats and tests will enable students to display their levels of understanding and competence through the specific mode(s) in which they are most competent. Some students perform well with essay tests, some on M-C tests, while others excel in doing projects, oral reports and presentations.

Before suggestions for constructing M-C items can be discussed, a familiarity with the basic parts of an M-C item is helpful. The stem is the main part of the item, that which precedes the choices or responses. The stem may be: (a) an incomplete sentence, (b) a question, or (c) a stated problem, graph or diagram. Students who are just gaining experience with M-C items are usually most successful with the question format. The possible answers, responses, or choices are also called foils or alternatives. All choices except the "correct" one are called distractors.

Major difficulties encountered in constructing M-C items include: (a) the development of a stem which is expressed clearly and without ambiguity, (b) the statement of an answer which cannot be refuted, and (c) the creation of options or distractors which are attractive to those who do not possess the knowledge or understanding necessary to recognize and select the correct answer.

The following suggestions are intended to identify potential problems in constructing multiple-choice items and to help solve some of them.

(1) The objective to be assessed by an item should be carefully reviewed before beginning to construct the item. If the item is to be useful, it must require students to perform the same behavior as stated in the objective. Fellow teachers can be most helpful in determining whether an item-to-objective match is accurate and appropriate. The "acid test" is, of course, the students. A brief question to a student--"What did you do to choose your answer?"--might be very informative.

(2) The stem should be phrased in simple and understandable language. The main criterion here is to use key words and phrases that are consistent with the level of instructional material and the background of the students. Solution of the item should depend on whether or not the students possess a command of the intended objective instead of an advanced reading comprehension or vocabulary. Unless it is essential for the objective tested, difficult technical vocabulary should be avoided. Overly complex sentences might be transformed into several separate sentences to enhance student comprehension of the problem. Irrelevant words or phrases which have no function (sometimes called window dressing) should be eliminated from the stem. Pictures, diagrams or tables of data are excellent ways to present some problems. These techniques, when appropriate, will also minimize the ever-present problem of overwhelming reading demands.

The stem should include all words that would otherwise be repeated in each of the responses. By including the common material in the stem, the differences among the responses are more obvious to the students.

Negative statements (especially double negatives) cause students difficulty and confusion. If a negative statement is the logical way of stating a particular problem or question, it is recommended that the negative word (no, not, except) be underlined and/or capitalized to draw student attention to the "reverse thinking" that is required. If a significant number of negatively stated items are to be included in a test, they might be grouped together in a separate section.

(3) The choice keyed "correct" must be unquestionably the correct response. If items become a matter of opinion, not achievement, they are of minimal value to the test. Colleagues can help provide this necessary check without involving too much of their time. A pair of teachers (or all teachers in a department) could routinely inspect each other's tests to make sure that the items have one and only one clearly correct response.

(4) Distractors should be constructed which are plausible and attractive to students who do not possess knowledge of the objective tested. All the responses must be grammatically consistent with the stem and parallel with one another in form, e.g., all action statements or all people. 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 offered, 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.

An article by Guttman and Schlesinger (33) offers several suggestions for constructing distractors systematically. For instance, with items that require students to perform numerical calculations, instead of choosing distractors at random or numbers near the correct answer, construct distractors based on the application of the wrong formula, the copying of a number appearing in the problem, or on other common conceptual or computational errors. Distractors can also be based on mistaken ideas or misconceptions commonly held among students. Many such misconceptions become apparent during class sessions, and Tamir (89) found that student responses to essay questions provided a fertile field from which to harvest plausible distractors for multiple-choice items. If these kinds of techniques are employed, considerably more information is obtained than merely whether the student got the item right or wrong. Through systematically constructed distractors, the thinking processes of the students can be diagnosed. Such feedback can also be helpful to the teacher concerned about the effectiveness of the instructional program and materials.

(5) 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%
2. less than 40%
3. more 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%
2. between 20% and 40%
3. between 40% and 60%
4. more than 60%

(6) A common problem in constructing M-C 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.

The responses "none of the above" and "all of the above" are overused. 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 1/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 an M-C item with several weak distractors.

(7) If some logical order exists among the responses, they should be arranged in that order to help the student in choosing. For example, when the responses are numbers, they should be arranged in ascending order. The following exception to this recommendation is noted for responses which include numbers between 1 and 5. Students may confuse the absolute value of the answer with the keyed number of the response.

| | | |
|---|---|---|
| POOR | 1. 1<br>2. 3 | 3. 4<br>4. 6 |
| BETTER | 1. 1<br>2. 6 | 3. 3<br>4. 4 |

(8) 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 more problems than they solve.

(9) Avoid patterns among 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 alternation 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.

**Measuring More than Facts**

Many of those who criticize science tests charge that these tests often measure only definitions, formulas and picayune bits of factual information. For some of the standardized and teacher-made tests currently in use, unfortunately, these criticisms are valid. Such recall or memory type items are much easier for both inexperienced and experienced people to write than are items measuring higher level objectives, so more low-level items get written and ultimately, more are included on tests. If one criterion for item selection is the correlation to items on an existing test (as it is for many national test publishing companies), items are collected which, over time, come to resemble one another in a variety of ways--including the level to be tested. Since items assessing higher abilities do not always correlate very highly with recall items, it is hard to "crack the starting lineup."

Pancella (71) analyzed the cognitive level of items found in commercially prepared biology examinations. Bloom's scheme of "Cognitive Levels" was used to categorize the items for 41 high school biology tests. The analysis of the items was validated by a "panel of 12 distinguished judges including four contributors to the Taxonomy." Across all these tests, 2689 items were determined to be distributed among the levels as follows:

| | |
|---|---|
| Knowledge | 71.80% |
| Comprehension | 15.17% |
| Application | 11.49% |
| Analysis | 1.37% |
| Synthesis | 0.04% |
| Evaluation | 0.04% |

Only 39 items--1.45% of the total--were above the application level. In five tests, 100% of the items tested only the knowledge level, while six other tests were composed of 90% or more knowledge items. The only tests which contained items above the application level were produced as a part of the BSCS project. From all the tests analyzed, only one synthesis and one evaluation item were identified--both from the Processes of Science Tests (POST) developed by BSCS (75). Pancella recommends that the POST be used as a model for teachers who wish to develop items measuring cognitive processes higher than knowledge.

As recommended by Pancella, the POST included several types of item presentations designed to assess outcomes above the recall level. Tabular or graphical presentation of data can be useful in such assessment. These data can allow students to demonstrate their skills in interpreting, analyzing and other higher order behaviors. A second exemplary technique is based on a description--verbal and/or pictorial--of an experimental situation. Based on this, the student can be questioned about conclusions, experimental variables, research design and a number of ideas central to scientific thinking.

Sets of predetermined responses have been used for assessing higher level outcomes. Burmeister (16) developed nineteen sets of keyed responses which relate to thirteen components of scientific investigations. One of the sets was the following:

A. Causes Hypothesis A to be rejected.
B. Causes Hypothesis B to be rejected.
C. Causes both hypotheses to be rejected.
D. Causes neither hypothesis to be rejected.

These keys can be applied to problem situations in all science areas with considerable success, although first attempts may prove difficult.

Another way to assess a particular set of knowledge without evoking responses on the recall level involves the use of hypothetical or fictionalized material. Students must tap more deeply into their understanding of a problem when it is posed in new or unfamiliar terms, even though the problem itself may be no more difficult than others they have encountered. As an example of this technique, the following item was developed by Sam Rotella (79). That a few chuckles may be generated during a science test is not perceived to be a drawback to using this kind of item, although using too many such items would certainly diminish their value and prove disruptive to the testing situation.

1. On Mars, orange hair is dominant to yellow hair and green skin is dominant to blue skin. With these facts and your own knowledge of genetics, solve the following problem.

   Marty--a homozygous tall, heterozygous green-skinne and heterozygous orange-haired, color-blind, skinny Martian--married Roda--a short, plump, blue-skinned, heterozygous orange-haired Martian who is a carrier of color blindness. They had one boy and one girl Martian.

   Describe their children in terms of these characteristics. Give all possible geno-& phenotypes of all possible children and the ratios you would expect.

Although designed for those with college-level biology classes, the CUEBS publication, Testing and Evaluation in the Biological Sciences (92), is an excellent source for illustrative items above the recall level. Although most of the nearly 1400 items are of the M-C type, some items use the essay and matching formats. Other item formats used in the CUEBS publication are called interpretative (a variation of the matching format), some using reading passages and others designed for open book tests.

Klopfer (50) presented exemplary items from various science fields which can aid in constructing items that measure achievement above the knowledge level. These items are grouped according to a category system he developed which had the following levels of "processes of scientific inquiry" in addition to "knowledge and comprehension":

- Observing and measuring
- Seeing a problem and seeking ways to solve it
- Interpreting data and formulating generalizations
- Building, testing, and revising a theoretical model
- Applying scientific knowledge and methods

Many of the suggestions for assessing objectives above the recall level have been described in terms of the "process as content." An article by Doran (20) summarized some of the attempts to measure these objectives at both the elementary and high school levels. Some of the tests utilized the presentation of stimuli by means of photographic slides and movie film, as well as by actual objects and materials. These last techniques minimize the reading demands on the students.

The assessment of objectives and outcomes at higher cognitive levels is an obviously difficult task requiring much creativity and perseverance. The time required to produce an item is much greater at the higher levels of

the Bloom cognitive taxonomy. Even though only a small percentage of these items may be relevant, their inclusion will do much to reinforce the importance of higher level behaviors in science courses.

## Mechanical Aspects of Test Construction

Much attention has been given to the skills necessary for developing and revising valid test items. Here the discussion will center on some of the important tasks involved in presenting the items to the students (by means of a test paper and directions); collecting responses (by means of answer sheets); and scoring the student responses. These are essential to the success of the test and can save the teacher time and effort. Several of the ideas discussed here are modified from those in the pamphlet, Improving the Classroom Test (40).

### Item Format

If teachers feel no need to number items, they may initially neglect to do so--only to regret it later. Certainly for reference to item analysis data and for student questioning, some means of identifying items is helpful.

The stems and options for items should not be broken up. For each item, all necessary materials--including graphs, data, reading passages and responses--should appear on the same page. When a test booklet is being used, it is sometimes admissible to have reference material on a page facing the item, but in many cases, items can be moved within the test to eliminate this problem.

Completion, true-false and matching items are very efficient in terms of the space they occupy on a test sheet. Several styles of multiple-choice item layout are illustrated below:

COMPACT LAYOUT: An example of an inclined plane is a:
(a) saw (b) hammer (c) car jack
(d) ramp

VERTICAL LAYOUT: The most commonly described effect that occurs as one's velocity approaches the speed of light is the:

a. dilation of elapsed time
b. decrease of mass of objects
c. increase of force of gravity
d. decrease of electrical attraction

COMPROMISE LAYOUT: The man who conceived most of the principles we now call the Theory of Relativity was:

a. Isaac Newton
b. Wernher Von Braun
c. Albert Einstein
d. Galileo

The compact layout uses the test paper most efficiently, but could be confusing to students if the responses are longer than one or two words. The vertical layout is necessary for long, complex responses to ensure that students can differentiate between the alternatives. For responses of just a few words, like the names above, the compromise layout is both effective and efficient.

Scoring Arrangements

The use of a separate answer sheet is especially advantageous for tests with three or more pages. Using an answer sheet eliminates the processes of making separate scoring stencils for each page, adding the scores for each page, and adding the page subtotals to get a total score. The need for an answer sheet system increases as the size of the student population increases. As errors resulting from fatigue and tedium are minimized and as efficiency is increased, the test may become more valid and reliable. Answer sheets could also allow the test booklets to be used more than once if the test could maintain validity with readministration.

If homemade answer sheets are used, the blanks should be arranged so that they are readable, convenient for students to use, large enough to write in, and efficient for teachers to score. Some teachers prefer student responses to be in blank spaces at the left margin while others prefer the right margin; in either case, the students must easily be able to tell which blank refers to which item. To avoid possible misinterpretations of student handwriting and erasures, it is not advisable to require students to write out answers, letters or numbers. Confusion often exists between "T" and "F" on True-False items, and among lower case letters on multiple-choice items. For M-C items, some teachers require the use of capital letters (A, B, C, D) or Arabic numerals (1, 2, 3, 4) which aren't quite as easy to confuse. But especially with younger students, the preferred method requires them to circle, underline, or check the letter or number corresponding to the response they have selected as correct.

Considering the wide use of machine scorable answer sheets, it is worthwhile to expose students to this system prior to the final exams or standardized tests that normally use this kind of answer sheet. With most of the machine scorable sheets, students must simply darken a space containing the number or letter corresponding to their choice of answer.

Distribution of Correct Responses

As mentioned before, a regular pattern of responses and a predominance of items with the same response number or letter should be avoided. It is worth taking a few minutes to check the distribution of correct answers and eliminate obvious sequences. Some teachers unconsciously write more true items than false ones, and others do the opposite. Similarly, some tend to consistently position their correct answers to M-C items in, for instance, the third position. Many students are alert to these nuances, and can use them in answering items of which they are uncertain. The frequency of correct answers for each response position should be checked and the distribution should be as even as possible.

Grouping and Arrangement of Items

In most instances, items of the same format should be grouped together, especially if unique directions are being used. Most people agree that it is best to arrange items within a test in terms of graduated difficulty. The first part of a test, especially, should include relatively easy items, giving students a chance for a good start and minimizing initial confusion or discouragement. If a test is set up in separate parts (e.g., based on item type or content), the first few items of each group should be the easier items.

Designating Credit Allowances

The amount of credit (or number of points) that may be earned on each section and item of a test should be made "perfectly clear" to the students so that they can allocate their time accordingly. This is especially true for essay questions and problem-solving items which involve mathematical computation or the "setup" of a problem solution. If partial credit is to be given for separate calculations or components of an item, students should be clearly and completely informed.

Directions for Answering Questions

Most students have become so accustomed to objective test items over their years in school that they seem to complete items almost ignoring

directions, if given. Nevertheless, a clear, specific set of instructions should be written for each set of items with a different format. The directions could be enclosed in a box, or written in a different size, style or color of printing in effort to encourage students to focus on what is being said. The use of sample items is very helpful in explaining new response procedures to students.

Correction for Guessing

What to do about guessing factors and chance scores has been discussed for years, but no agreement has yet been reached. The guessing factor is dependent on the item format and several other factors. A commonly used formula is the following:

$S = R - \frac{W}{k}$ Where S = the calculated guessing score
R = the number of right answers
W = the number of wrong answers
k = the number of options in each item

This formula assumes that all wrong answers are chosen primarily by guessing--a questionable presumption. Most guesses are made by eliminating implausible distractors and detecting grammatical inconsistencies, response patterns and other clues--a combination of factors for which no formula can adequately account. In general, corrections for guessing should not be used, but if a "guessing correction" is to be made, students should be explicitly informed so that they may adjust their test-taking approach accordingly, responding only to those items for which they are quite confident of the correct answer.

Even when a guessing correction is not used, students should be reminded of the scoring procedure (e.g., that their score is the total number of items they answer correctly). In these situations, students should be encouraged to attempt all items, choosing their best guess when they aren't sure of the answer.

Allowing Choice of Items

Some teachers allow students to choose among essay questions on a test. On some multiple-choice tests, students may choose among several groups of items. The New York State Regents Examinations in science commonly follow this latter practice to accommodate the variety of topics stressed in some schools and omitted in others. The problem with such procedures is that unless each group of items is uniformly difficult, students are actually taking different tests.

In most cases, students should be required to complete all items in a given test. If choice is allowed, include directions that are clear and precise and ensure that the choosing will not unduly affect the validity and reliability of the test.

Printing and Duplicating

Very short and simple tests may be written on the blackboard, but the vast majority of tests should be duplicated so that students may have their own copies of the test. Most schools have some kind of machine that will produce copies of adequate quality. If items are written or printed by hand, be sure they are neat and legible. The size of print should be comparable to that in the reading materials the student uses. Many typewriters have a special setting which produces a darker image for the preparation of various kinds of stencils. Use generous margins and spaces between items. And in all cases, proofread the stencils before producing the copies. Also before the test is duplicated, the answer key should be prepared. This can help in spotting errors in the items that may have been missed previously.

With experience, judgement improves about how many items of various types can be completed by certain students in a given period of time. As a rule of thumb for beginners, consider that students can complete 1 1/2 M-C items per minute and 2 to 3 T-F items per minute. There are many factors which influence these guidelines, however, so individual determinations must be made about the number and kind of items to use on any given test. The nature of the response desired is the biggest determiner of time required. Simple recall items take little time, whereas more complex application and analysis items demand and consume considerably more student time.

# CHAPTER THREE

# Assessing Affective Outcomes in Science

**Introduction**

Although the emphasis on cognitive objectives and abilities persists, many of the more recently developed school science programs stress goals within the affective domain. These goals include not only the identification of attitudes, interests, and values, but also their subsequent cultivation. According to Shulman and Tamir (84),

> We are entering an era when we will be asked to acknowledge the importance of affect, imagination, intuition and attitude as outcomes of science instruction that are at least as important as their cognitive counterparts.

Yet, Klopfer (50) cites the "paucity of informed, analytical discussion of affective behaviors in science education until now." These outcomes are not meant to replace cognitive and psychomotor objectives, but rather to assist in the integration and amplification of all learning. Interaction among the three domains was recognized by Bloom and his co-workers (13) in their first handbook. Interchange is most noticeable in the affective-cognitive relationship, but also occurs between the other pairs. Klopfer (50) concludes that "It is already amply clear. . . that a student's attitudes and interests are always associated with cognitive elements." And Harbeck (35) asserts that the affective domain is central to all learning and evaluation processes:

> Awareness initiates learning. Willingness to respond is the basis for psychomotor response, and value systems provide the motivation for continued learning and for most of an individual's overt behavior.
>
> The affective domain cannot be ignored, regardless of the difficulties encountered. Teachers do show values and students do develop values. Often what the teacher does speaks louder than what he says. It seems safer to continue to consciously work on evaluation than to leave it to chance and hope for the best.

The major reason for the scarcity of adequate measures in the affective domain is the lack of clear definitions of affective variables. This situation is further complicated by the broad range of constructs included within this domain--those mentioned before, plus adjustment, awareness, appreciation, feeling, and orientation. Although these variables have much in common, they may be more clearly defined and isolated as the psychometric properties of the measurement devices improve. Recent work by Fraser (25) provides hope that these goals can be attained.

Another area of difficulty involves the type of response format and scoring system for affective measures. Since "keyed" responses are generally inconsistent with this domain, responses must be summarized by a frequency distribution or by a range of replies. Different statistical procedures must be applied to these kinds of response data than are used with cognitive data.

For the science teacher who is just beginning to measure affective outcomes, an excellent reference is Behavioral Objectives in the Affective Domain by Eiss and Harbeck (22). This NSTA publication is cited at numerous points in this chapter.

## Conceptions of the Science Affective Domain

The second handbook of the Taxonomy of Educational Objectives by Krathwohl, Bloom, and Masia (53) deals with the affective domain and has become the overriding authority in the field since its publication in 1964. The five levels of the affective taxonomy established by Krathwohl and his associates are 1) Receiving; 2) Responding; 3) Valuing; 4) Organization; and 5) Characterization by a Value or Value Complex. As the authors point out, these terms overlap in meaning with more commonly used but less exactly defined affective terms like adjustment, value, attitude, appreciation and

interest. Because the words are new to the contexts in which they are used, the Krathwohl levels sometimes present problems for those working with them for the first time. To aid in comprehending the domain and its levels, Gronlund (31) has assembled two tables which encompass the definitions of the five Krathwohl levels; illustrative general objectives of instruction; and examples of behavioral terms appropriate to each level.

Nay and Crocker (67) analyzed the affective attributes of scientists and produced a comprehensive inventory including interests, operational adjustments, attitudes or intellectual adjustments, appreciations, and values and/or beliefs. According to the authors, "this scheme is not meant to be hierarchical. No serious attempt was made to separate attributes into more or less important ones, or to arrange them sequentially." After some difficulty in differentiating between attitudes and adjustments, they interpret the latter to mean adjustments to the requirement of the dynamics of science, and describe these as being "operational" in nature. Those that are intellectual in nature are commonly referred to as "attitudes." Each attribute is stated in positive terms, but this does not mean that every scientist or every science student is expected to act in that way at all times. This five-stage inventory could be of substantial help to a science teacher attempting to measure science affective outcomes and therefore is included on the following page.

Many of the discussions of science objectives and outcomes in the affective domain are based on varying concepts of the "scientific attitude." The range of subjectivity involved in formulating these concepts is indicative of the difficulty encountered when trying to define--or even outline--the boundaries of the affective domain. The line between "the willingness to make value judgments" and compulsive judgmentality, or between "critical-mindedness" and dour skepticism, or between "humility" and self-effacement, is sometimes as fine and difficult to draw as the line between thinking and feeling, apprehension and appreciation, ideas and ideals, even fact and opinion. Where does one end and another begin? And what about the grey areas in between?

As these questions suggest, there is no clear-cut definition either of the affective domain or of what constitutes the scientific attitude. This ambiguity, however, does not abrogate attempts to clarify some of the dimensions of both. For instance, Haney (34) summarized that:

"INVENTORY OF THE AFFECTIVE ATTRIBUTES OF SCIENTISTS"
REMOVED DUE TO COPYRIGHT RESTRICTIONS

> Attitudes have emotional content and vary in intensity and generality according to the range of objects or situations over which they apply. For the most part, attitudes are learned and are difficult to distinguish from such affective attributes of personality as interests, appreciations, likes, dislikes, opinion, values, ideals and character traits.

Haney claimed that "to be scientific, means that one has such attitudes as curiosity, rationality, suspended judgment, open-mindedness, critical-mindedness, objectivity, honesty and humility." In addition to describing and explaining each term, Haney suggested several procedures to aid teachers in the fostering of these attitudes.

Moore and Sutman (66) developed and validated a Scientific Attitude Inventory which was based on both intellectual and emotional attitudes, the former focusing on some knowledge about the object of the attitude, and the latter relating to a feeling or emotional reaction to the object. The following six statements, all stated positively, were used to form the inventory and were scored on a four point Likert-type scale. The first three statements emphasize the intellectual attitudes and the last three the emotional attitudes:

1. The laws and/or theories of science are approximations of truth and are subject to change.

2. Observation of natural phenomena is the basis of scientific explanation. Science is limited in that it can only answer questions about natural phenomena and sometimes it is not able to do that.

3. To operate in a scientific manner, one must display such traits as intellectual honesty, dependence upon objective observation of natural events, and willingness to alter one's position on the basis of sufficient evidence.

4. Science is an idea-generating activity. It is devoted to providing explanations of natural phenomena. Its value lies in its theoretical aspects.

5. Progress in science requires public support in this age of science, therefore, the public should be made aware of the nature of science and what it attempts to do. The public can understand science and it ultimately benefits from scientific work.

6. Being a scientist or working in a job requiring scientific knowledge and thinking would be a very interesting and rewarding life's work. I would like to do scientific work.

While there is no "accepted" list of components of the "scientific attitude," several common elements emerge from the many different schemes that have been developed. This commonality, however, is more suggestive than definitive, and awareness of the variable and volatile aspects of the science affective domain will help prevent falling into the trap of establishing hard and fast rules for affective objectives and outcomes. The goal of affective measurement, after all, is not to "rule out," but to "rule in"--to assist in the identification of student attitudes and abilities and then to encourage the application of that which has been identified to a specific context, like the subject of science.

Eliciting the affective response and focusing it in terms of science are the tasks of science affective assessment. To do these things effectively, educators and examiners must be aware of their own biases and preconceptions and must take care that they do not intrude upon a fair and "open" assessment of attitudes that do exist but which are often not afforded an avenue of expression.

Belt (10) has noted that "the image of science and scientists that we elicit. . .is highly dependent on the method of questioning used and on the type of question asked." Much space is allocated here to the way questions are asked--the various questioning techniques and strategies that may be employed in affective assessment--but that is only half the story. The other half is the questions themselves--both those that are asked and those that remain unasked. Not only do unasked questions go unanswered, they also may never be asked another time. The effect of silence is often to subtly undermine the significance of the subject altogether.

Practical tips can be given on the "how" of affective assessment, but the "what" is just as important and much more difficult to prescribe. One guideline for determining what kinds of questions to ask is the criterion of "inclusivity" (as opposed to exclusivity)--asking constructive, open questions on which to build further questions or responsive actions. If students are asked if they like science, they may say "no," and then the teacher is stuck. But if students are asked what they do like, the creative teacher can take their responses and put them to work where they seem to apply most--in the laboratory, at the computer terminal, in the field or the library. The possibilities are as limitless as they are challenging, to students and teachers alike. Perhaps the most that can be done here to outline the

potential power of the "what" is to encourage an active consideration of the problem and to urge a sensitivity to its complexity and, at the same time, its essentiality.

## Techniques for Assessing Affective Outcomes

Just as a variety of methods may be used to measure and assess cognitive outcomes, there are also several ways of collecting student responses relating to affective objectives. In the pages that follow, some of these affective measurement techniques are described and examples of them are given. One technique is not necessarily superior to another, and no single technique will be adequate for measuring the wide range of affective behaviors pertinent to any given subject area. All of these methods, however, are highly adaptable; the challenge lies in selecting the means most appropriate to the end, or outcome, being measured.

### Semantic Differential

This technique was originated by Charles E. Osgood and colleagues (70), as part of a research study of meaning. It has become a widely used technique for measuring people's responses to a variety of words and ideas. The test developer selects a set of relevant bipolar adjectives or adjective phrases which defines a range of meaning with respect to the object being evaluated. This object may be almost anything--a subject area, an idea, an animate or inanimate being--but care must be taken to ensure that the bipolar terms chosen actually relate to the object to which they are to be applied, and that the two terms used are, in fact, opposites. Often these bipolar terms are classified into dimensions of potency, activity and evaluation. The following pairs of terms have been categorized as fitting these dimensions:

| EVALUATIVE DIMENSIONS | | POTENCY DIMENSIONS | | ACTIVITY DIMENSIONS | |
|---|---|---|---|---|---|
| Good | Bad | Large | Small | Active | Passive |
| Optimistic | Pessimistic | Potent | Impotent | Fast | Slow |
| Beneficial | Harmful | Sharp | Dull | Thick | Thin |
| Clean | Dirty | Strong | Weak | Moving | Still |
| Valuable | Worthless | Heavy | Light | Exciting | Dull |
| Helpful | Harmful | Complex | Simple | Intentional | Unintentional |
| Wise | Foolish | Hard | Soft | | |
| Useful | Useless | | | | |
| Important | Unimportant | | | | |

Usually a seven-point scale is arranged between the bipolar terms (X and Y) with the scaled positions defined as follows:

1. Extremely X
2. Quite X
3. Slightly X
4. Neither X nor Y, Equal X and Y
5. Slightly Y
6. Quite Y
7. Extremely Y

EXAMPLE

(X) School (Y)
(Object)

Good_____Bad
Useless_____Useful
Exciting_____Dull

(People have used more than seven scaled positions, but have found that little new information is obtained and that more positions merely confuse the respondents.) The bipolar terms should be arranged so that no regular distribution of the "positive" terms exists. The respondent checks the space on the continuum between the two words that best describes his feeling about the object. This general technique has been used in many ways, tailored to specific situations or needs. The following example is excerpted from Eiss and Harbeck (22):

SCIENCE IS

whee! . . . . . yetch!
theoretical . . . . . practical
inconvenient . . . . . convenient
complex . . . . . simple
wide . . . . . narrow
easy . . . . . troublesome
unnecessary . . . . . basic
dull . . . . . emotional
efficient . . . . . inefficient
universal . . . . . limited
outgoing . . . . . ingrown
broadly interpretive . . . . . dogmatic
imaginative . . . . . unimaginative
interesting . . . . . uninteresting
objective . . . . . subjective
clear . . . . . fuzzy
useful . . . . . harmful
good . . . . . bad
exciting . . . . . boring

To obtain valuable bases of comparison, responses to the same set of bipolar terms (though differently arranged) could be gathered using different school subjects such as social studies, math, music, art, physical education and English. A good example of this comparative use of the Semantic Differential (SD) scale is the following item adapted from the "Subject Area

Preference" portion of the IOX collection, Attitude Toward School (K-12) (7). The same six bipolar word pairs are used with each subject, but in different orders and with reversed polarities, so students can't easily fall into response sets.

SCIENCE IS

| | | |
|---|---|---|
| Meaningful: | ___:___:___:___:___:___:___: | Meaningless |
| Bad: | ___:___:___:___:___:___:___: | Good |
| Useful: | ___:___:___:___:___:___:___: | Useless |
| Confusing: | ___:___:___:___:___:___:___: | Clear |
| Unimportant: | ___:___:___:___:___:___:___: | Important |
| Simple: | ___:___:___:___:___:___:___: | Complex |

To interpret student responses, it is helpful to summarize class responses by preparing either a profile or a frequency distribution. By multiplying the number of student responses for each position between the bipolar terms by an arbitrary point value (1-7), and then dividing by the number of students, a "weighted mean" is obtained. (Although the point values are arbitrary, they should be assigned in a pattern consistent with the "positive" or "negative" aspect of the terms used, no matter how those terms are arranged on the scale. In the example above, for instance, if "Meaningful" is valued at 1 and "Meaningless" at 7, then "Bad" should be valued at 7 and "Good" at 1.) Given the following data for 50 students, the "weighted mean" is obtained as follows:

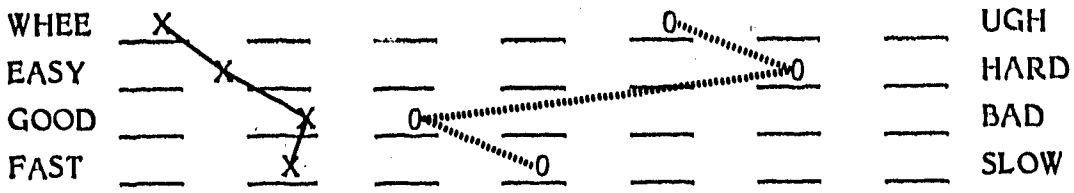
| SCIENCE IS | Point Value (V) | Frequency (F) | VxF |
|---|---|---|---|
| Extremely Meaningful ____ | 1 | 2 | 2 |
| Quite Meaningful ____ | 2 | 18 | 36 |
| Slightly Meaningful ____ | 3 | 14 | 42 |
| Neither Meaningful Nor Meaningless ____ | 4 | 5 | 20 |
| Slightly Meaningless ____ | 5 | 8 | 40 |
| Quite Meaningless ____ | 6 | 0 | 0 |
| Extremely Meaningless ____ | 7 | 3 | 21 |
| | | Total Students 50 | Weighted Sum 161 |

$$\text{Weighted Mean} = \frac{\text{Weighted Sum}}{\text{Total Students}} = \frac{161}{50} = 3.22$$

Although a wealth of information is lost when a series of numbers is reduced to one, simplification is sometimes necessary. Once such weighted

means are obtained, profiles can be produced for a variety of purposes. Separate profiles may be developed for different classes or sex groups. The following chart shows how one profile might appear:

SCIENCE IS




The placement of the marks on each line is based on weighted means as previously discussed. The solid line could represent the "pre" measure (before a science course) and the dotted line could be the "post" measure. If that is the case, several inferences may be made about what happened during the year. Profiles which allow comparisons among several groups of students or among object terms can be very useful in evaluating curriculum and teaching.

Likert Scale

This technique, developed by Renis Likert (56), is one of the most easily constructed and widely used scales for measuring attitudes. Scale developers need only construct declarative sentences, stated in either negative or positive terms, that are related to the topic being evaluated. Students respond to each statement by marking a position on a five-point continuum, usually comprised of the following terms: Strongly Agree, Agree, Uncertain or Neutral, Disagree, Strongly Disagree. Statements must be meaningful and interesting to the students. If many students are skipping items or consistently choosing the middle response, the statements are not functioning as intended. Possible "shorthand" symbols used with Likert scales include the following:

| | | | |
|---|---|---|---|
| Strongly Agree | SA | AA | 5 |
| Agree | A | A | 4 |
| Uncertain, Neutral | U | N | 3 |
| Disagree | D | D | 2 |
| Strongly Disagree | SD | DD | 1 |

The following Likert scale items were included in Allen's (3) instrument to assess the "Attitudes of Certain High School Seniors Toward Science and Scientific Careers."

INSTRUCTIONS: Please give your reactions to the following list of statements regarding science, scientists, and scientific careers. Work rapidly. Record your first impression--the feeling that comes to mind as you read the item.

Draw a circle around AA if you completely agree with the item.
Draw a circle around A if you are in partial agreement.
Draw a circle around N if you are neutral.
Draw a circle around D if you partially disagree.
Draw a circle around DD if you totally disagree.

| | | |
|---|---|---|
| AA A N D DD | (1) | The development of new ideas is the scientist's greatest source of satisfaction. |
| AA A N D DD | (2) | Scientific investigations are undertaken as a means of achieving economic gains. |
| AA A N D DD | (3) | Modern science is too complicated for the average citizen to understand and appreciate. |
| AA A N D DD | (4) | The working scientist believes that nature is orderly rather than disorderly. |
| AA A N D DD | (5) | Scientists are willing to change their ideas and beliefs when confronted by new evidence. |
| AA A N D DD | (6) | Science and its inventions have caused more harm than good. |

After obtaining student responses to each item, a variety of techniques can be used to help interpret the results. As suggested for the Semantic Differential technique, a weighted mean could be computed for each statement, or a frequency distribution could be made of the number of students responding to each position for each statement. If, as was suggested earlier, a value of five is assigned to each SA response, then agreement with statements considered to be negative (items 2, 3, and 6 above) would skew the scoring in a direction opposite than intended. By reversing the scoring for such negative statements, higher scores could then be interpreted as being in greater accord with the conceptualized position. If the intent is to produce an overall score reflecting a "positive attitude toward science," the scoring for items 2, 3, and 6 would have to be the reverse of that for item 1 and other "positive" statements.

Forced Choice Items

In this format, students must choose one of several responses to a question or statement. The similarities to the Likert scale are obvious, although the item construction is different and a wider variety of responses is possible. The following examples from Shoresman (83) and Airasian (2) respectively demonstrate this technique:

DIRECTIONS: Each sentence below has a blank space in the middle. Following each sentence are five ways you can fill the blank. After you read the sentence carefully, choose the one answer which is MOST like the way you really FEEL. Choose ONLY ONE answer for each sentence. Remember, there are no right or wrong answers to any of these sentences. When you have decided which answer is most like the way you feel, CIRCLE the letter in front of your choice.

1. I like reading about a great writer_____reading about scientific discoveries.
   a. a lot more than
   b. a little more than
   c. just as much as
   d. a little less than
   e. a lot less than

DIRECTIONS: Below is a list of physical science and science-related activities. Rate these on a scale from 3 to 1 (as follows):

3 If you would like to attempt such an activity
2 If you are indifferent about attempting such an activity
1 If you would not like to attempt such an activity

You are to assume that you have ample time to attempt any activity which may interest you.

3 2 1 Discuss in a chemistry class the nature of chemical bonding.

3 2 1 Include some science books in your general reading program.

Other items of this style may use the following response systems:

| | | |
|---|---|---|
| 1--Definitely | 1--Frequently or Regularly | 1--Very Little or None |
| 2--Probably | 2--Occasionally | 2--Sometimes, but less than once a week |
| 3--Probably Not | 3--Hardly Ever | 3--About once a week |
| 4--Definitely Not | 4--Never | 4--Twice or more a week |

Personal Discussions and Interviews

Often a direct approach brings excellent results. Mager (60) tells a story that shows the value of direct questioning. The U.S. Army was trying to determine which recruits would be most efficient working at a base in Alaska near the Arctic Circle. The Army administered various psychological tests, collected all sorts of data about the physical and physiological functioning of the recruits' bodies, and conducted in-depth psychiatric interviews. All these data were no more effective in predicting the success of recruits than was simply asking the question, "Do you like cold weather?" The moral of the story is that simple ways of collecting affective data should not be overlooked.

Using personal discussion and interviews was suggested by Eiss and Harbeck (22) as an excellent way of determining student attitudes and values. They cautioned that personal questioning can result in answers students think the teacher wants. This is especially true with questions like, "How well do you like science?" and "Why do you think science is important?" Eiss and Harbeck suggested the following questions to help identify student values without giving value cues:

a. What subjects do you like most?
b. What do you do in your spare time?
c. What hobbies do you have?
d. Do you like to visit museums?
e. Have you made a career choice? What is it?

Knowing students and their interests and backgrounds can aid in the selection and adaptation of questions that will be effective.

Student Reports and Term Papers

While these techniques do not directly collect affective information, a student's choice of topics for reports, projects, or papers may give some insight into the student's interests and values. Eiss and Harbeck point out that even a report which is a "dry recital of facts gleaned from source books and encyclopedias" may be useful, as it may "indicate a total lack of interest in science and should motivate the instructor to try to develop an instructional program that will be more meaningful to the student." Scientific overtones or implications in a student's report, paper, or project for an English, math, or social studies class may indicate an interest in science. Similarly, science teachers can share with colleagues the writing skills, math competence, and knowledge of social studies topics that are displayed in science classes.

Subjective Test Questions

Eiss and Harbeck suggested using subjective test questions that require students to exercise value judgments. They included the following questions which first ask the student to make a choice and then to give a rationale for the decision made.

a. The town council has been caught in a budget squeeze between the need for a new sewage disposal system for your community and the need for improved medical services at the local hospital. You have been invited as a citizen to visit a council meeting and make recommendations for action. What would you recommend and what reasons would you give to support your decision?

b. Suppose that the science club, of which you are a member, is planning its year's activities. What activities would you suggest for the club and what reasons would you give to encourage others to support your selection?

Students should be graded not on their choices, but on the reasons used within their answers and the supporting evidence they offer. Questions that reflect current and controversial concerns within the school and community are ideal for this purpose.

Checklists

These instruments are used in an attempt to implement the idea that the things students do (or fail to do, or refrain from doing) are the best indicators of their interests and values. One way to collect data on student behavior is to develop a checklist on which to indicate the occurrence and frequency of certain behaviors. This system can be more focused, comprehensive, and objective than what has been called the "anecdotal" reporting system in which descriptions of student behavior are recorded in a narrative style the way in which the behavior occurs, with no attempt to organize or otherwise structure the information recorded.

A checklist may be constructed by listing behaviors consistent with the goals of the class, the science program, or the field of science itself. Students under observation cannot be expected to display every behavior each day, week or month. By using the same checklist over a period of time, however, a teacher can begin to make a balanced assessment of student behaviors and may make inferences about student attitudes toward the teacher, the class, or science in general. Eiss and Harbeck provide examples of the kinds of behavior that may be observed and recorded using a checklist format:

## VERBAL BEHAVIORS

**Argues:**
- Advocates desirable actions
- Defends desirable actions
- Criticizes plans and suggestions

**Asks:**
- Inquires for further information
- Examines others' ideas by further questioning

**Explains:**
- What others have said
- Personal ideas
- Principles and theories
- Reports on a science topic

**Reads:**
- Science magazines
- Science books
- Science articles in the daily or weekly press

## NON-VERBAL BEHAVIORS

**Participates:**
- Joins science clubs
- Participates actively in science clubs

**Contributes:**
- Time to science projects
- Money to science projects
- Time and money to agencies attempting to improve man's environment

**Purchases:**
- Scientific reading materials
- Science equipment

**Borrows:**
- Science books
- Science equipment

**Selects:**
- Discriminates between useful materials and "gadgets"
- Signs up for advanced science courses
- A science-related career

**Visits:**
- Science centers
- Hospitals, health centers
- Research laboratories

**Assists:**
- In laboratory preparation and operation

**Eats:**
- Nutritionally balanced meals

**Repairs:**
- And adjusts scie[illegible]e equipment

**Builds:**
- Science-related equipment

**Works:**
- Part-time in science-related job

Multiple-Choice Items

The M-C format is widely used for assessing cognitive outcomes. Additionally, it is used in some affective instruments, such as the Test on Understanding Science developed by Cooley and Klopfer (19). Recently Kozlow and Nay (52) used multiple-choice items on their Test on Scientific Attitude (TOSA) which assesses components of both cognition and intent in attitudinal behavior. They argued that:

> The multiple-choice item includes a stem describing a situation relevant to a given attitude and distractors describing different courses of action. This is consonant with the position taken in this study that an individual's attitude can be inferred from his endorsement of certain courses of action relevant to the attitude, object or situation.

The following two items from the TOSA illustrate the use of the multiple-choice item in testing components of both cognition and intent, respectively.

1. Scientists recognize that a scientific theory
   - A. Should not be changed when it is based on a large amount of data.
   - B. May have to be changed to keep up with a rapidly changing world.
   - C. May have to be changed when new observations are made.
   - D. Should not be changed when it explains what happens to nature.

38. If you come across a scientific item which goes against your common sense, which one of the following would you be inclined to do?
   - A. Disregard the scientific idea because it is better to rely on common sense.
   - B. Disregard common sense because it is not as reliable as scientific study.
   - C. Do an experiment to see whether or not the common sense is superior to the scientific idea.
   - D. Try to produce a compromise between the scientific idea and common sense.

Q-Sort

Although the Q-Sort has been available since the mid 1930's, it has not yet been used widely in the classroom. According to Humphreys and Townsend (38), "Q-Sorts are believed to produce a more honest assessment of attitudes than most questionnaire measures." To use this technique, the investigator prepares a number of words or phrases describing a trait or

subject in many ways, from highly positive to highly negative. Each word or phrase is recorded on an individual card. In most cases, the student is asked to sort the cards into an order most reflective of his or her response to the subject, forming a sequence from most agreeable to most disagreeable, most representative to least representative, etc. In the Humphreys and Townsend research, students were asked to sort 50 adjectives describing their "ability to achieve." Among the 50 words used were capable, confident, observant, successful, incompetent, lazy, careless and awkward. The distribution of the placement of the cards among a group of students may be charted if statistical data are desired. In terms of classroom advantages, the Q-Sort technique is highly adaptable and relatively unobtrusive to administer.

Projective Techniques

Widely used in psychology, these techniques have been explored for use in science education. Lowery (58) developed an open-ended attitude instrument composed of three interdependent projective techniques--the Word Association Test, the Lawrence Lowery Apperception Test (LLAT) and the Sentence Completion Test. It was "designed to delve beneath the surface of superficial answers in an attempt to uncover hidden attitudes which are not revealed through ordinary methods." A modified version of Lowery's attitude instrument was used by Gallagher and Korth (27) in the Ohio Test Every Senior Project.

In the Word Association Test, several selected words--such as science, experiment, and scientists--are placed at random among other words having no specific science orientation. The list of words is read aloud to the student one at a time, and the student is asked "to respond as rapidly as possible with the first three words that come into his mind" when he hears the stimulus word.

In the LLAT, the student is asked to interpret a drawing which depicts a specific theme but in an inspecific or open-ended way. The three drawings prepared for this test are as follows: "The picture for the first theme (science) shows the child reading the headline of a newspaper on a newsstand. The headline states, 'NEW SCIENCE DISCOVERY.' The second theme (process) pictures a child looking at a science experiment. The third shows the child meeting a scientist." Each child is shown one drawing at a time; asked to "make a story to suit himself;" and told that there are no right or wrong answers.

In the Sentence Completion Test, each student is asked to finish nine sentences, three for each of the themes (science; process; scientist). In each group of three sentences, one is designed to be positive, one negative, and one neutral. Sample sentences are:

The field of science is_____.
Most people like science whenever it_____.
One thing tha uts some people against science is _____.

This novel technique is especially appropriate and adaptable to students at the elementary and middle/junior high school levels, as Hofman (36) has demonstrated in her "Assessment of Eight Year Old Children's Attitudes Toward Science."

Thurstone Scale

Another technique for affective assessment which originated in the field of psychology is the Thurstone Scale. This kind of scale consists of a series of statements, each of which has been constructed to fit into a particular position on a response continuum ranging from disagree to agree (or unfavorable to favorable) with respect to whatever topic is to be assessed.

Billeh and Zakhariades (11) developed a Scientific Attitude Scale (SAS) using this technique. Each statement was sorted by a panel of 45 judges "into one of eleven piles, where No. 11 indicates the most favorable feelings toward the psychological object, No. 1 indicates the most unfavorable feelings, and No. 6 is determined as a neutral point expressing [illegible] favorable nor unfavorable feelings." The SAS consisted of 36 statements which formed a hierarchy of views from the most favorable to the most unfavorable with respect to science. Of the SAS items, the item considered to be the most favorable was the following:

Newly discovered ideas should be reported unchanged even if they contradict existing ones.

The statement labeled "most unfavorable" by the panel of experts was:

It is worthless to listen to a new idea unless all people accept it.

The "scores" for students who respond to the SAS (with Disagree or Agree) are calculated based on their responses and the "scale value" of each item. This scale value is determined by the median of the judgments from the selected panel.

## Developing a School Assessment Program

Science education has traditionally emphasized the cognitive objectives and outcomes, and there has been considerable reluctance on the part of teachers to involve themselves in the assessment of affective outcomes. Some of the reasons for the avoidance of these objectives have been cited by Birnie (12) as involving:

1. A general feeling that trying to develop attitudes and values in students is akin to indoctrination and brainwashing;
2. The inadequacy of available methods and materials designed for use in this domain;
3. The general dearth of evaluation instruments and techniques in the area of science affective measurement.

Other reasons include the ill-definition of the domain itself and of objectives designed to guide affective behavior (particularly disturbing to those who work with the exacting terms of science as a rule); and, more practically speaking, a serious deterrent to the introduction of affective assessment is the great amount of effort required to initiate and establish a thoroughgoing program.

Nevertheless, interest in the affective domain is growing and science education appears to be slowly accepting responsibility in this area. Building on existing foundations--course outlines, teaching materials, evaluation instruments--is a good way to begin. From the outset, teachers, students, administrators, and parents should be involved in the challenge of planning and assessing the affective dimension of school science programs. Feedback--the opportunity for it, the encouragement of it, and the responsiveness to it--is the most important feature in any such planning stage. Although affective assessments may be administered in conjunction with cognitive tests or lab skill exams, ultimately it may be determined that they are best conceptualized and implemented independently. If an independent program is undertaken, literature relevant to program development has emerged over the past decade and should be made available to those interested and involved in the planning process.

Two articles focusing specifically on affective goals in science are one by Birnie (12) and one by Klopfer (49). In terms of the actual evaluation task, reviewing the research of Belt (10), Aiken and Aiken (1), Pearl (73), and Kozlow and Nay (52) may save "reinventing the wheel" in affective measurement while benefiting from the tested applications of others.

Concerning the variety of assessment formats available and their features of simplicity, sensitivity, and interpretability, parts of the following may be helpful: Gephart, Ingle, and Marshall (29), Stanley and Hopkins (87), Tuckman (94), Nunnally (68), and Bloom, Hastings, and Madaus (14).

To review some of the things to keep in mind when approaching affective assessment, remember that the kinds of questions asked are as significant as the method of questioning chosen. Feedback--both in the development of the assessment program and in the actual collection of data--is essential. The data collected do not always have to assume the form of a student grade, but rather may be better used to evaluate teaching effectiveness, to supplement instructional materials, or to reassess the goals of the entire science program. Whatever the end use of the data, the purpose of the evaluation should be accurately and honestly communicated to the students before the assessment is made, and then the data should be used for the stated purpose and only that purpose.

The following suggestions were among those presented by Eiss and Harbeck (22) to those interested in beginning to establish affective objectives and evaluate affective outcomes in their science programs:

- A good program evolves; it is not created. Provide plenty of opportunity for revision and change as you proceed.
- Use a variety of evaluation instruments. No single method of observing affective behaviors will be adequate.
- Allow for individual differences.
- Trends are more important than absolute attainments. Look for trends and encourage students who show favorable changes.
- Be honest and open-minded.
- Be prepared for change.
- Look for leaders. They are key individuals who influence the others profoundly.
- Experiment with new ideas. Experiments don't always succeed, so be prepared for failures.
- Try taping a class session, either with a sound tape recorder or a video-tape. Analyze the tape to see if the lesson was teacher-centered or student-centered. Ask questions like:
  a. What percent of the time was the teacher talking?
  b. To what extent did students have the opportunity to discuss their problems and ideas?
  c. How often was theory presented as fact?
  d. How many student suggestions or ideas were received and acted upon?
  e. How authoritarian was the teacher?
  f. Is the atmosphere in the classroom conducive to the free exchange of ideas?
  g. Who held the center of attention? Who contributed most of the ideas presented?

# CHAPTER FOUR

# Assessing the Outcomes of Science Laboratory Activity

## Introduction

No one doubts that scientists do lab work. According to Thomas (93) this has become "one of the fundamental tenets of our dogma." But how the laboratory can best be used as part of the school science program is still an unanswered question. Shulman and Tamir (84) assert that with the advent of the new curricula which stress the processes of science and emphasize the development of higher cognitive skills, the laboratory has "acquired a central role, not as a means for demonstration and confirmation, but rather as the core of the science learning process." The implementation of this view, however, has been difficult.

In the laboratory, students can learn to perform particular laboratory skills and procedures; formulate hypotheses and interpret data; and develop interest in and attitudes about the processes and purposes of science. These outcomes relate to the psychomotor, cognitive, and affective domains. Although manipulative outcomes are experienced predominantly in the laboratory setting, few rationalize labs solely on the basis of skill development. Kreitler and Kreitler (54) believe that the unique contributions of the experiment to science instruction are its ability to provide a basic means for developing conceptual thinking and imagination (by "evaluating the raised alternatives as possible solutions"), and its fostering of scientific practices in the classroom. Other objectives and outcomes of the laboratory

experience have been reviewed by Fuhrman, et al. (26).

Based on their analyses of the BSCS, PSSC, and other laboratory handbooks, Lunetta and Tamir (59) concluded that:

> Students are commonly asked to make observations and measurements, record results, manipulate apparatus, and draw conclusions. On the other hand, they are given few opportunities to discuss sources of experimental error, to hypothesize and propose tests, or to design and then actually perform an experiment.
>
> Thus, in spite of the curriculum reform of the last 20 years, students still commonly work as technicians, following explicit instructions and concentrating on the development of lower level skills.

## The Learning Domains and the Science Laboratory

The development of the cognitive and affective domains and their application into instruction, curriculum, and research have far outdistanced that of the psychomotor domain. The psychomotor (also called perceptual-motor or motor manipulative) domain was initially conceived by Bloom and associates (13) as the third major area in which educational objectives could be categorized.

According to Simpson (85), the psychomotor domain is relevant to education in general as well as to specialized areas. Singer (86) elaborates that "psychomotor activity is associated with military tasks, agricultural duties, industrial, professional, technical, and vocational skills, driving demands, music, art, and dance works, as well as physical education, sports and recreation endeavors." Surprisingly, however, science is absent from both Singer's and Simpson's lists of fields with psychomotor components. It may be inferred that science education is perceived to emphasize primarily cognitive goals despite claims about the centrality of the laboratory to science instruction.

One difficulty in developing schemes for organizing science laboratory objectives may be a result of the inherent interaction of all three domains. This interlocking nature was described by Jewett and colleagues (42):

> . . . no learning experience can be classified exclusively in any one of three domains. It is obvious that cognition has a motor base and that experiences resulting in significant affective outcomes are devoid of neither cognitive nor motor aspects. Similarly, all objectives classified in the motor domain probably have some degree of involvement in both the cognitive and affective domains. Thus, classification is a matter of emphasis.

The interaction of the several domains of learning outcomes is similarly expressed by Moore (65):

> Although some recognized educational goals, such as typing skills and piano playing, may lie obviously and almost exclusively within the perceptual-motor domain, it is evident that, in this view, the perceptual-motor domain encompasses the domains of both cognition and affect; and spans a far greater developmental range than either. Perhaps herein lies the chief contribution that its detailed consideration may make to education.

The overlap is clear when examining the two following descriptions of the components or outcomes of science laboratory activities.

| Eglen and Kempa (21) | Jeffrey (41) |
|---|---|
| 1. methodical procedure | 1. vocabulary competence |
| 2. experimental technique | 2. observational competence |
| 3. manual dexterity | 3. investigative competence |
| 4. orderliness | 4. reporting competence |
| | 5. manipulative competence |
| | 6. laboratory discipline |

Cognitive, affective, and psychomotor elements are included in both lists.

## Evaluating Science Laboratory Outcomes

Just as many different kinds of objectives may be served by means of laboratory activity, so have several different styles of assessing student lab skills been employed. These include (1) paper-and-pencil test items, (2) checklists or rating scales which require a teacher to observe a student performing a given operation, (3) lab reports, and (4) the laboratory practical examination.

(1) The test item formats parallel those used in the cognitive domain; for lab-related objectives, multiple-choice, matching and completion items are those most frequently used. Matching items are usually of the type in which students are presented with laboratory apparatus (or photographic slides of the apparatus) and then asked to identify each piece of equipment from a list provided. Examples of the kinds of apparatus that might be presented range from the basic--beakers, pipettes, test tube holders--to the more specialized--deflagration spoons, retorts, water aspirators, etc. Completion items for the laboratory may also focus on the equipment in use there, but such items can additionally survey experimental procedures and outcomes.

There are many examples of multiple-choice items which require students to use or interpret data or which present information in graphical form. The following item constructed by Ruda (80) relates to a typical chemistry problem with data used for identifying "unknowns."

You were given an unknown pure substance and your data table after many tests on the sample appears as follows:

| TEST | RESULTS |
|---|---|
| Boiling point | 81°C |
| Freezing point | 5.6°C |
| Density | 0.88g/ml |
| Solubility in water | Insoluble |
| Solubility in ethanol | Very soluble |

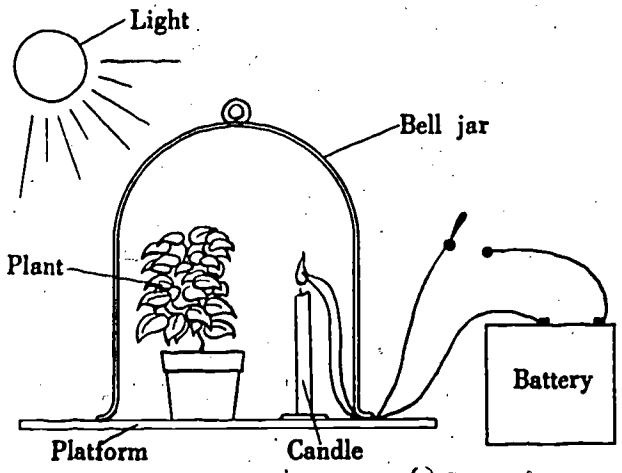
Using the table below, your unknown most closely resembles:
(a) Oleic acid (b) Cyclohexane (c) Benzene (d) Chloroform

Table

| | BP | FP | Density | Solubility $H_2O$ | Solubility Alcohol |
|---|---|---|---|---|---|
| Oleic acid | 285 | 16 | .891 | insol | ∞ |
| Cyclohexane | 80.7 | 6.5 | .779 | insol | ∞ |
| Benzene | 80.1 | 5.5 | .879 | 0.07 | ∞ |
| Chloroform | 61.2 | -63.5 | 1.489 | 0.82 | ∞ |

Since considerable space is required for such items and much time is required of students to read and assimilate prior to choosing an answer, clusters of items are commonly developed around one presentation or description. The cluster of items from Korth's (51) Life Science Process Test appearing opposite illustrates this technique.

Questions 33-36 relate to the following experiment:






82

"LIFE SCIENCE PROCESS TEST" REMOVED DUE TO
COPYRIGHT RESTRICTIONS

(2) One of the first checklists to be used to evaluate science laboratory behavior was Tyler's (95) checklist for assessing students' skill with a microscope. In it the teacher notes the sequence of the student's actions, checking skills in which the student needs further training, and listing noticeable characteristics of the student's behavior. The following section excerpted from the Tyler checklist illustrates its format and detail.

Tyler Microscope Checklist

The student's goal is to find a specimen present in a culture. The teacher's goal is to see whether the student is able to operate a microscope so that the specimen is located. The student is provided with all the necessary materials and the teacher observes his actions, numbering them in the order of their occurrence. In addition to actions directly related to finding the specimen, other actions are listed concerning areas that require further training, the student's behavior, and the mount itself.

| | STUDENT'S ACTIONS | SEQUENCE OF ACTIONS |
|---|---|---|
| a. | Takes slides | 1 |
| b. | Wipes slide with lens paper | 2 |
| c. | Wipes slide with cloth | |
| d. | Wipes slide with finger | |
| e. | Moves bottle of culture along the table | |
| f. | Places drop or two of culture on slide | 3 |
| g. | Adds more culture | |
| h. | Adds few drops of water | |
| i. | Hunts for cover glasses | 4 |
| j. | Wipes cover glass with lens paper | 5 |
| k. | Wipes cover glass with cloth | |
| l. | Wipes cover with finger | |
| m. | Adjusts cover with finger | |
| n. | Wipes off surplus fluid | |
| o. | Places slide on stage | 6 |
| p. | Looks through eyepiece with right eye | |
| q. | Looks through eyepiece with left eye | 7 |
| r. | Turns to objective of lowest power | 9 |
| s. | Turns to low power objective | |
| t. | Turns to high power objective | |
| u. | Holds one eye closed | 8 |
| v. | Looks for light | |
| w. | Adjusts concave mirror | |
| x. | Adjusts plane mirror | |
| y. | Adjusts diaphragm | |
| z. | Does not touch diaphragm | 10 |
| aa. | With eye at eyepiece turns down coarse adjustment | 11 |
| ab. | Breaks cover glass | 12 |
| ac. | Breaks slide | |

The limitations of the checklist mode are obvious: it is restricted to a one-to-one situation and is often not generalizable beyond the specific operations observed. It may, however, be a very valuable technique to use in ascertaining student lab skills related to a particular lab, unit or course.

As an aid to assessing students' manipulative skills in practical chemistry, Eglen and Kempa (21) developed three types of assessment schedules: an open-ended schedule, an intermediate schedule, and a checklist mode. A portion of their checklist is shown below:

ASSESSING MANIPULATIVE SKILLS IN PRACTICAL CHEMISTRY

III. Checklist mode

Section A--Dissolution of the solid in water

| | | YES | NO |
|---|---|---|---|
| a.1. | Were the beaker and the stirring rod initially washed with distilled water? | ☐ | ☐ |
| a.2. | After the transfer of the solid into the beaker, was the weighing bottle rinsed out with water? | ☐ | ☐ |
| a.3. | Was the addition of water to the solid in the beaker done carefully, i.e., down the side of the beaker? | ☐ | ☐ |
| a.4. | Was the solution stirred until all the solid had dissolved? | ☐ | ☐ |
| b.1. | Was the beaker containing the solution adequately supported during stirring? | ☐ | ☐ |
| b.2. | Was the stirring action itself safe and satisfactory? | ☐ | ☐ |
| b.3. | Were all operations carried out in a manner which did not conflict with the quantitative nature of the exercise? | ☐ | ☐ |

Although specific to a solution procedure, the organization and style of these questions could be adapted and applied to many different laboratory settings and situations.

To aid the observation process, checklists or scales could be developed for each unit of content or cluster of lab skills. The degree of proficiency or skill with which each student accomplishes any given task may be indicated by using a set of numbers representing several levels of skills; e.g., 1 = Inadequate, 2 = Adequate, 3 = Superior. A more elaborate example of an

observation procedure utilizing levels of skills is the following, which was suggested by Hofstein, Lunetta, and Giddings (37).

Observational Assessment Criteria

| Skill Area | Criteria | Score |
|---|---|---|
| A. Planning and Design | 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 |

Appearing opposite is a card on which the instructor might record observations of lab skills using these assessment criteria for a number of separate experiments and for a variety of skill areas.

(3) Another traditional way to assess laboratory performance involves examining the students' written lab reports, through which considerable insight into student skills and/or deficiencies may be gleaned. Many of the procedures and criteria used for evaluating by checklist may be directly applied to reviewing lab reports and can serve to avert subjective consideration of factors extraneous to laboratory performance. Hofstein, Lunetta, and Giddings (37) have enumerated some of these factors and warn against the exclusive reliance on lab reports in assessing laboratory activity, but they assert that "On the other hand, used sensitively, lab reports can serve as an appropriate mechanism for stimulating student dialogue and interaction as well as for providing one source of evaluative data."

"SAMPLE SKILLS ASSESSMENT CARD" REMOVED DUE TO COPYRIGHT RESTRICTIONS

(4) As an aid to the objective assessment of student responses to laboratory performance tests, or "lab practicals," Tamir (89) developed a "Practical Tests Assessment Inventory" (PTAI). This inventory could be useful for constructing checklists and for evaluating lab reports as well as for assessing laboratory practical examinations. The PTAI consists of 21 categories of laboratory skills, similar to those within the scheme developed by Lunetta and Tamir (59) mentioned earlier. However, the PTAI extends their work by specifying behaviors within the different skill levels while simplifying the point values accordingly assigned. The following is an example from the PTAI illustrating possible skill areas, student behaviors, and their respective point values:

MAKING GRAPHS

| | | |
|---|---|---|
| a. | Drawing the Graph | |
| | Adequate and perfect drawing | 5 |
| | No or inadequate title | 4 |
| | Inadequate scaling and relation of x and y axes | 3 |
| | Inadequate connection between points of the graph | 2 |
| | Combination of at least two of the above | 1 |
| b. | Recording of Variables | |
| | Dependent variable on y axis and independent variable on x axis | 6 |
| | Independent variable on y axis and dependent variable on x axis | 5 |
| | Inappropriate recording of variable names and units | 4 |
| | No recording of the variable names and units | 3 |
| | Confusing the variables on the axes | 2 |
| | Combination of at least 2 from the above | 1 |

Each category has different specific skills appropriate to that category. The number of possible points varies with the category. Tamir described in detail how to use the PTAI for assessing a particular item of a lab performance test.

Some of the advantages of the laboratory performance test noted by Kruglak (55) and Wall, et al. (96) include its utility in measuring skills (like psychomotor responses) that are not easily assessable elsewhere; its low frequency of answer "leak over" from one set of students to the next; and its potential to increase motivation, improve outcome measurements, and expand the grade range. Necessary to the success of lab practicals are adequate time allotment (typically, 2 hours); durable, reliable equipment; and complete, detailed instructions.

Examples of experiments that have been used in the BSCS program to assess practical laboratory skills include measuring the rates of photosynthesis, human respiration, grasshopper respiration, and yeast fermentation; the alternation of activity in daphnia; and water relations of plant tissue. Student information, questions, and the Examiner's Guide for one such experiment are presented here as they appeared in the BSCS instrument developed by Tamir and Glassman (90).

SET-UP:

Problem 1. Measuring the rate of photosynthesis

On the table are three beakers filled with water. In each is an inverted funnel containing several sprigs of fresh elodea. On the funnels are calibrated test tubes. The first set-up is in the direct light provided by a 100-watt lamp. The second set-up is about one meter distance from the lamp. The third set-up is completely concealed under a heavy paper cylinder. There are also two liter bottles containing a solution $NaHCO_3$. (If the set-ups are arranged about one hour before the students arrive, there is a clearly discernible difference in gas level in the test tubes.)

TO THE STUDENT:

Problem 1.

1. Examine the rate of photosynthesis of the three set-ups in front of you. Write the results.
2. What is the control in this experiment?
3. How would you explain the results? Indicate the major processes occurring in each of the set-ups.
4. What is the gas that collects in each of the test tubes? How can you test this?
5. Why did we use a water plant in this experiment? (elodea)

EXAMINER'S GUIDE:

Testing Procedure & Evaluation of Responses

Problem 1.

1. The student will have to measure the rate of photosynthesis by observing the accumulation of gas in each of the three t.t.s for 10 minutes.
2. This experiment has no control. If the student did not mention this he would lose five points.
3. In his explanation the student is expected to indicate the processes going on and the reasons for the observed differences.
4. The student will suggest how the test is to be carried out--but he is not asked to perform the actual test.
5. The dependence of the method of measurement on the type of plant selected is to be explained.

According to the scoring guide, the examiners observed students' self-reliance and manipulative skills. Scores on observation, investigation, communication and reasoning were based on students' written answers. The

relatively high reliabilities obtained with this practical exam prompted Tamir and Glassman to suggest that teachers and schools incorporate such examinations into their existing procedures.

## Illustrative Assessment Techniques

Since it is often helpful to see examples of different assessment techniques applied to various learning objectives and outcomes, some illustrations for the different stages of the Lunetta and Tamir (59) model are provided in the following pages. The examples, of course, are merely suggestive of the range of assessment styles and applications possible in the classroom laboratory. The model appears opposite.

### Planning and Design

Student abilities in this skill area may be assessed by a series of separate test items, a checklist for monitoring student plans, or a laboratory practical examination.

Talesnick (88) has developed an excellent laboratory achievement test comprised of many different individual problems for students to encounter. For each problem--e.g., "correctly identify the contents of a series of vials containing 'unknown' liquids"--the students must first design an experiment. The following illustrates this phase of the Talesnick test.

Problem

> The labels from five laboratory containers came off the containers and were mixed up. The labels, listed in alphabetical order, are barium hydroxide, calcium carbonate, citric acid, sodium chloride and sugar.
>
> Individual samples of the five materials are contained in the vials labelled A, B, C, D and E.
>
> Using only the materials in the "SPECIAL LAB KIT," design an experiment to correctly identify the contents of the five vials.
>
> The design must be written in detail on the Scoring Guide in Section A--Experimental Design.
>
> Do NOT proceed with the actual experimental work until the examiner has checked and approved the experimental design that you have suggested. NOTE: You are also provided with a kit of Standard Laboratory glassware and hardware.

"LABORATORY STRUCTURE AND TASK ANALYSIS INVENTORY"
REMOVED DUE TO COPYRIGHT RESTRICTIONS

Talesnick allows the students 15 minutes for this phase. The instructor then examines each student's design and rates it on the following scale:

| | |
|---|---|
| Good and workable | 5 |
| Faulty (does not require an initial clue) | 3 |
| Faulty (requires an initial clue) | 2 |
| None | 0 |

If the student's design is not sufficient to start performing the laboratory work, the instructor provides appropriate clues.

The value of Talesnick's instrument lies in its requirement that students produce an original experimental design--a high-level objective--within carefully articulated and reasonably limited boundaries. Virtually all the sub-category skills of planning and design will be brought into play as the students outline the experimental procedure, but the task is saved from being overwhelming by its clear definition. Another valuable aspect of the instrument--not illustrated here--is the continuity it provides among the remaining task categories (performance, analysis, etc.) as the students are later required to carry out and follow up the experiment they themselves initially planned.

The following item, developed by Ruda (80) as part of a chemistry laboratory practical exam, is an example of a paper-and-pencil item assessing, in a more limited way, yet another dimension of the planning and design capability:

> Consider the numbered steps listed below. They are all steps you would carry out to determine the concentration of an unknown acid by base titration. The correct order to carry out these steps would be:
>
> 1. Add base until indicator changes
> 2. Add indicator
> 3. Calculate concentration of acid
> 4. Determine volume of base used
> 5. Fill buret with standard base
> 6. Measure known amount of acid into titration vessel
>
> (a) 6,2,5,1,4,3
> (b) 6,2,1,5,3,4
> (c) 5,6,1,2,3,4
> (d) 6,5,4,1,2,3

The following test items are from the ERIE Science Process Test (97) and the Processes of Science Test (75) respectively. They illustrate formats that may be useful for assessing some of the other dimensions of competency in experimental planning, like defining the problems investigated and confronting questions of experimental design.

A tire company wants to know if they will get as much mileage from a new type of tire as from their usual tire. Which one of the following variables would it be most important to control in an experiment?

1. The time of day the test is made
2. The number of miles traveled by each type of tire
3. The physical condition of the driver
4. The weather conditions
5. The weight of the car used

Which of these experimental procedures would serve best to determine the effectiveness of inoculating sheep against anthrax disease?

(A) Expose 50 sheep to anthrax and then inoculate all of them
(B) Inoculate 25 out of 50 sheep and then expose all 50 to anthrax
(C) Inoculate 50 sheep and then expose all of them to anthrax
(D) Inoculate 25 out of 50 sheep and then expose only the 25 inoculated sheep to anthrax

According to Anderson (6), the planning and design or "thinking" aspects of school science laboratory programs have been minimized to favor an emphasis on the "manipulative" aspects. In terms of the role the laboratory plays in what scientists actually do, trend is hardly representative. An effective school science laboratory program should integrate the variety of skill objectives involved in experimental procedures and assess the outcomes accordingly.

Performance

This manipulative phase of laboratory activity has received much attention in laboratory manuals and assessment procedures. The Tyler checklist and BSCS Practical Lab Exam illustrated earlier are examples of appropriate techniques.

The finals exam for the New York State Regents Earth Science syllabus (76) includes a five-task performance test. Students proceed sequentially to each lab station where they perform a particular measurement task. The following is the information provided to the teacher to help regulate the equipment used and the scoring procedure.

TASK NO. 3: VOLUME MEASUREMENT

Materials
- Graduated cylinder (100 ml)
- Water supply
- Irregularly shaped, nonporous, nonsoluble mineral specimen of sufficiently small size to fit easily into the graduated cylinder
  NOTE: The specimen MUST be nonporous and nonsoluble.

Preparation
- Select a sufficient number of appropriate sized mineral samples to meet your class needs.
- Code each sample.
- Measure and record the volume of each sample.
- Have a source of water at each station.

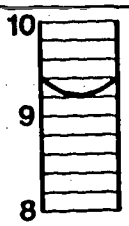
Scoring
- A student response within plus or minus 1.0 ml of the teacher-determined volume of a given sample will receive 2 points.
- A response within plus or minus 2.0 ml will receive 1 point.
- A response range greater than plus or minus 2.0 ml will receive no credit.

The following items from Ruda (80) relate to other sub-categories of performance competency.

A. Go to station 6. Using the pH paper provided, measure the pH of one of the solutions in the beakers. Record the pH and code number of the solution used.

Code Number_____pH_____

B. Examine the picture of the graduated cylinder at the right. Record the volume in the cylinder to the nearest 0.1 ml.




C. Go to station 3 and, using a pipet, measure 10.0 ml of acid into a 125 ml Erlenmeyer flask. Add approximately 25 ml of distilled $H_2O$ and add 3-4 drops of the phenolphthalein indicator. Titrate, with the base in the buret, to the pink endpoint. Record the volume of base needed to reach the endpoint and the code number of the acid solution used.

Code Number_____Volume Required_____

D. Go to station 7. Add 10 drops of liquids A, B, and C, one to each test tube. Now add 10 drops of liquid X to each test tube. Record your observations and rank the liquids A, B, and C in terms of their reactivity (low to high) with X.

E. A sample of $KClO_3$ was decomposed to yield KCl and $O_2$. From the following data determine the number of grams of $O_2$ in the sample and the percentage of $O_2$ in the sample. SHOW ALL WORK.

| | |
|---|---|
| Mass of crucible and cover | 36.48g |
| Mass of crucible, cover and sample | 39.98g |
| Mass of crucible, cover and sample after reaction | 39.92g |

Mass of $O_2$ in the sample_____

% of $O_2$ in the sample_____

Analysis and Interpretation

This stage is another component of the "thinking" aspect of science laboratory behavior. The following series of questions is addressed to students as part of a problem evaluating the chemical reactions that take place when unidentified substances are introduced into a potato. It is excerpted from the BSCS Laboratory Exam by Tamir and Glassman (90).

Problem 6.

(Steps 1-3 are directions to the student to prepare the potato.)

4. Observe what occurs in twenty minutes and write your observation.
5. What is the explanation for this phenomenon?
6. What do you think are the substances put in each hole? What is the basis for your hypothesis?
7. How would you test your hypothesis? (Hint: Look at the materials on the table. Call the examiner and show him your plan.)
3. Make a table of the tests and their results. What do you now think are substances A and B?
   - Taste each of the substances. What do you think they are?
   - Show the examiner what you have written in this paragraph.
9. Look again at the experiment. How much time has passed since the beginning? Can you notice any change since your previous observation?
10. Based on your observations and your tests, do you still think that your explanation in item 5 is the best one? If not, suggest a new one.
11. Had you taken a sugar beet instead of a potato, would you get the same results? Explain.

Several of the Lunetta and Tamir sub-categories are brought into play throughout this series of questions. For example, within question #8 students are requested to "make a table of results" (3.1a), and within #5 they are asked to formulate an explanation or generalization (3.5).

Another approach involves sampling specific behaviors with specific items applied in a number of different laboratory situations over the duration of the science course. As examples of such specifically oriented items, the two problems below require students to make a table and a graph of data derived from experimentation, representing sub-categories 3.1a and 3.1b respectively.

1. Given the following information, make a table displaying this data according to increasing height:

   Sam--120 cm, 35 kg; Burt--150 cm, 45 kg;
   Ron--195 cm, 85 kg; Al--165 cm, 60 kg;
   Jim--180 cm, 80 kg; Bob--135 cm, 40 kg.

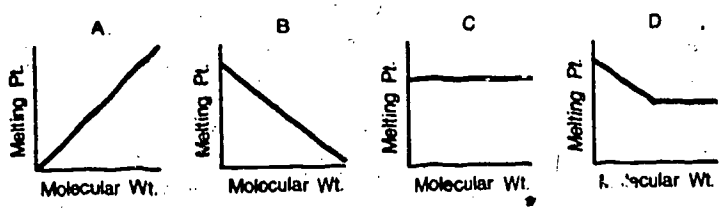
2. Graph the following data.

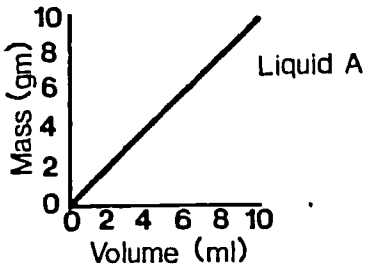
| Temperature (°C) | Solubility of Sugar in Water (g/100 ml) |
|---|---|
| 0 | 180 |
| 20 | 200 |
| 40 | 240 |
| 60 | 290 |
| 80 | 360 |
| 100 | 490 |

The sub-categories 3.2a and 3.2b relate to the determination of qualitative and quantitative relationships. The first item below requires students to choose the graphical relationship that best represents a table of data, whereas the second item requires students to determine a quantitative relationship between mass and volume--specifically, density. Both items are from Ruda's (80) exam.

Which graph below best represents the relationship between molecular weight and melting point as indicated by the following data:

| Substance | Molecular Weight | Melting Point (°C) |
|---|---|---|
| A | 32 | -97.8 |
| B | 46 | -117.8 |
| C | 60 | -127.0 |
| D | 74 | -136.3 |






What is the density of liquid A?_____

Determination of experimental accuracy (3.3) is illustrated in Ruda's exam by the following item which typifies error calculation in science laboratory work.

Composition of a Hydrate Experiment

From the data given below calculate the experimental percent of $H_2O$ in the hydrate of barium chloride. Calculate the percent error in your determination. SHOW ALL WORK.

| | |
|---|---|
| Mass crucible and cover | 29.81g |
| Mass crucible, cover and sample | 35.10g |
| Mass crucible, cover and sample after heating | 34.28g |

Theoretical percentage = 14.8%

The last three items included in this section exemplify possible ways of assessing the higher level tasks of formulating generalizations or interpretations (3.5), explaining relationships (3.6), and formulating new questions or hypotheses (3.7). The response sets illustrating these modes are from items by Korth (51), Ruda (80) and Korth, in the order in which they appear.

The best interpretation of the data from these two studies is that:

1. The first investigation does not support the second.
2. All ants respond in the same way to sunlight.
3. Ant activity increases as the temperature increases.
4. The general pattern of ant activity is the same in each investigation.

Using only the graph shown below, you would conclude that the:

(a) solubility of gases decreases with decreasing temperature.
(b) solubility of gases decreases with increasing temperature.
(c) solubility of oxygen gas increases with decreasing temperature.
(d) solubility of oxygen gas decreases with decreasing temperature.

The hypothesis best supported by the data from the second study is that:

1. Unknown forces control ant behavior.
2. Ant activity is about the same regardless of the temperature.
3. Ant activity is greater at a morning temperature of $70^{o}$ than at an afternoon temperature of $70^{o}$.
4. Ant activity seems to be related to the temperature.

Application

As the name implies, this stage involves students in taking the hypotheses, results, and experimental techniques gleaned from one investigation and utilizing them in another experimental or problem situation. The following test item from Ruda (80) is typical of the kind of outcome (prediction--sub-category 4.1) that might be commonly expected to emerge from an investigation.

In a study of the relationship between the volume and temperature of a sample gas, the following data were obtained:

| Temperature $^{o}C$ | Volume (ml) |
|---|---|
| 30 | 20 |
| 40 | 22 |
| 50 | 25 |
| 60 | 27 |
| 70 | 29 |
| 80 | 31 |

What would you predict the volume to be at $75^{o}C$?

(a) 28.0 ml (b) 29.5 ml (c) 30.0 ml (d) 30.5 ml

Examples of questions which measure application behaviors within a lab practical setting are the following from the BSCS Lab Practical exam by Tamir and Glassman (90).

Excerpted from Problem #1

9. What are your conclusions from all the experiments that you did?
10. Write down a hypothesis based on the results of the experiments you performed.
11. Describe, in short, how you can test your hypothesis experimentally.

Excerpted from Problem #5

7. To continue your work, choose one of the yeast suspensions. Design two different experiments to slow down the rate of fermentation. For each experiment, write down a hypothesis.
8. Perform the experiments and record the results.
9. What was the control in these experiments?
10. What is the role of fermentation in the life of the yeast?
11. What is the gas created during this process? How can you prove this?
12. What are the conclusions of your experiments?

Question #10 from the first problem is an illustration of an item assessing sub-category 4.2--requiring the formulation of a non-abstract hypothesis. Similarly, the series of questions in Problem #5 requires students to extend an experimental technique to a new problem (4.3).

A fifth category was added to the Lunetta and Tamir model after their work was published. After articulating the category, Hofstein, Lunetta, and Giddings (37) described a student exhibiting exemplary "Responsibility, Initiative, and Work Habits" in this way:

> Self-reliant, able to work with little supervision. Willing to tackle problems. Can work as part of a team as well as on own. Safety conscious. Willing to help running of laboratory if asked. Consistent and perseveres. Tackles practical work with enthusiasm.

These behaviors appear to be most amenable to some form of teacher observation. The following section from Allen's (4) checklist illustrates some specific behaviors that are relevant to this category.

Section B

| | | |
|---|---|---|
| 1. | Brings questions and/or activities to class. | |
| 2. | Can work in a group. | |
| 3. | Can work independently. | |
| 4. | Persists with an area of interest. | |
| 5. | Can say "I don't know." | |
| 6. | Displays initiative. | |
| 7. | Displays skill. | |
| 8. | Asks for help when needed. | |
| 9. | Refuses help when appropriate. | |
| 10. | Asks relevant questions. | |
| 11. | Suggests a way of solving a problem. | |
| 12. | Challenges ideas, that is, is skeptical. | |
| 13. | Contributes a fact. | |
| 14. | Contributes an explanation. | |
| 15. | Works steadily. | |
| 16. | Gets excited about science. | |

The appropriate stress on skills development in science programs is still a moot question. Each science teacher will differ in the emphasis she/he gives to the students' equipment manipulation and laboratory technique. Research into the several aspects of science laboratory objectives is woefully lacking. There are currently no universally accepted criteria for describing a student's science laboratory skills. According to Klopfer (50), a major problem is:

> . . . to find ways of developing much more detailed and precise specifications than have heretofore been attempted of the behaviors that the student is to attain. These specifications would also delineate the prerequisite behaviors leading to the desired criterion behavior, so that the student who has not attained mastery may be given soundly based guidance.

Compared to the testing of cognitive objectives, little has been accomplished in the assessment of laboratory-related objectives. After examining the evaluation programs of science curriculum projects, Grobman (30) concluded that "There has been little testing which requires actual performance in a real situation, or in a simulated situation which approaches reality." This condition was recognized by Tamir and Glassman (90) even in the BSCS curricula where lab-centered activities figure as a significant part of the program. Grobman (30) recognized that "testing is difficult and expensive, yet since the long run primary aims of projects generally involve doing something rather than writing about something, this is an area which should not be neglected in evaluation of curricula."

Further study and analysis of science laboratory behaviors will have a positive impact on curriculum, instruction and evaluation. Inquiry into this concern should stimulate more philosophical and empirical investigations.

# CHAPTER FIVE

# Item and Test Analysis

**Introduction**

A test is no better than its constituent items and, in many cases, a course grade is little more than an accumulation of test scores. Regardless whether the assessment objective centers in measuring a cognitive grasp of concept, an affective code of conduct, or the coordination of skill and thought required in a laboratory setting, the test scores and course grades that teachers issue students represent a responsibility that both must share.

As with the teaching/learning dynamic, the dynamics of test-giving and -taking are difficult to break into their component parts. Teachers fashion the tests they give partly according to the limits and license afforded by the class. At the same time, students adjust their test-taking strategies from class to class depending on their careful appraisal of which teacher is likely to ask what. The responsibility for some parts of the testing dynamic, however, rests solely with the ones giving the tests.

Item and test analysis are the chores that teachers must tend to in order to ensure that their part of the testing and grading process is as valid and fair as possible. Analyses of option utility, response pattern, item performance, test validity and reliability are but a few of the details concerning the teacher after the test has been written and administered. Fortunately, much of this information can be generated by hand computations, with modern calculators, or increasingly, through computer programs.

## Item Analysis

The success with which suggestions for item construction have been incorporated may be gauged in many ways, some of which are more exacting than others. Two commonly employed parameters of item performance are indices of difficulty and discrimination.

### Difficulty

The most widely used parameter--the item difficulty index--is commonly defined as the proportion of a given sample choosing the response keyed "correct." As defined, it really should be called an "ease index." The label "difficulty index" has been widely used for years, however, and can be viewed as an arbitrary convention. The item difficulty value is often expressed in percentage terms (e.g., 78%) or the decimal equivalent (e.g., .78 or simply 78). It is often called the "P value," based on its being a proportion or percentage comprised of those who correctly answer an item. The calculation required to obtain an item difficulty value involves a division of the number of students choosing the right answer--R--by the total number of students in the sample--N. The formula for difficulty is often expressed as $P = R/N$.

During a class review of a test, a teacher could obtain a measure of item difficulty by asking students to raise their hands if they got the item right. A more exact method of obtaining calculations of item difficulty involves having students indicate their responses on machine-scorable answer sheets, having the answer sheets scored, and submitting the student response data to a computerized test analysis program which will calculate difficulty indices plus a host of other parameters to be discussed later.

A commonly used criterion for item difficulty is an extension of the "statistically magical" 50% level for test mean scores. If a test mean should be at the 50% level to be maximally discriminating, so must the difficulty of the items. From this standard emerges an "optimum value" for item difficulty (and test mean) which is halfway between the random guessing score and 100%. For two-option items this value is calculated to be 75%, which is halfway between 100% and the guessing level for two options--50%. For three-option items, the optimum value is 67%; for four-option items, 63%; and 60% for five-option items. Since it is unrealistic to expect to construct a test in which every item meets this optimum value, some have suggested using a range of values for "acceptable" item difficulty. One such

suggested range is 40% to 60%, an expansion of parameters that retains the optimum 50% level as its focus. A variety of other ranges has been suggested, including the 30% to 90% range which eliminates only the very easy items (above 90%) and the very hard (below 30%). The rationale for excluding both the very hard and very easy items is that they do not contribute much to the test's discriminability.

A third criterion for difficulty is a distribution of values which combines items of moderate difficulty with items of extreme difficulty and ease. One example of this criterion is:

| Difficulty | Range of Values | Percentage of Items |
|---|---|---|
| very easy | .85 - 1.00 | 15% |
| moderately easy | .60 - .85 | 35% |
| moderately difficult | .35 - .60 | 35% |
| very difficult | .00 - .35 | 15% |

The system employed should match the intent of the test, the nature of the students, and the level of instruction.

A course developed in accordance with a philosophy of "mastery learning" would necessitate a different set of criteria than would a norm-referenced course. Teachers should be cognizant of the difficulty level of items they construct and use. Knowledge of an item's difficulty value helps a teacher to determine whether to use, revise or discard certain items in future tests. Item difficulty information may also be used in constructing separate tests or sub-tests corresponding to grades "A," "B," or "C." A "C" test, for instance, could be composed of mostly easy items graduating to mostly difficult items.

Discrimination

The second test item parameter--the discrimination index--is sometimes called the validity index or the item power index. It is commonly defined as a measure of how well an item differentiates between the "high" students and the "low" students. The categories "high" and "low" are usually determined by means of an "internal criterion": the total score on the test of which any given item is a part. An example of an "external criterion" would be the students' scores on another test or some achievement battery. The score obtained from either criterion system is used to categorize students into "high" and "low" groups. The simplest method of categorization involves ordering total scores from a test selected as the criterion, and

then labeling the students with scores above the median as "high" and those below as "low." This is called the High and Low Half system. Others have suggested using the High and Low Third, Quarter, or other proportion. In these cases, the middle group of students is not used in the calculation. The resulting values from all these calculations are quite similar, and the simplest procedure--the High and Low Half--is recommended. If there are several students achieving the median score, they can either be eliminated from the computation or randomly assigned to the High and Low groups until both groups have the same number of students.

If N = the total number of students in the sample, H = the number of students in the high group and L = the number in the low group choosing the correct answer, the most common calculation for discrimination is:

$$D = \frac{(H - L)}{N/2}$$

One of the main attributes of this formula is that it yields values which fall between the limits of +1.00 and -1.00, a familiar range. The following chart is most helpful for hand calculation of difficulty and discrimination indices of items. It assumes that each student's responses have been scored and totaled. High and Low group assignations can then be made. The number of students from each group choosing each of the responses (correct choices and distractors) may then be tallied.

1. Which animal has the simplest nervous system?

A. Rattlesnake
*B. Hydra
C. Owl
D. Turtle

| Item #1 | Group Size | Responses A | B | C | D | Omit | Difficulty P=R/N | Discrimination $D=\frac{(H-L)}{N/2}$ |
|---|---|---|---|---|---|---|---|---|
| High | 25 | 3 | 20 | 1 | 1 | 0 | 30/50= | $\frac{(20-10)}{50/2}$= |
| Low | 25 | 8 | 10 | 5 | 2 | 0 | .60 | +.40 |

This item was administered to a sample of 50 students, 25 in the High group and 25 in the Low group. The correct choice--B--was chosen by 20 students in the High group and 10 in the Low group, or 30 students in all. The difficulty of the item is calculated to be 60%, a moderately easy item. The discrimination index is calculated to be +.40--a very adequate value. Each of the distractors (A, C, and D) was chosen by more students in the Low group than in the High group. According to these data, each distractor appears to be performing well.

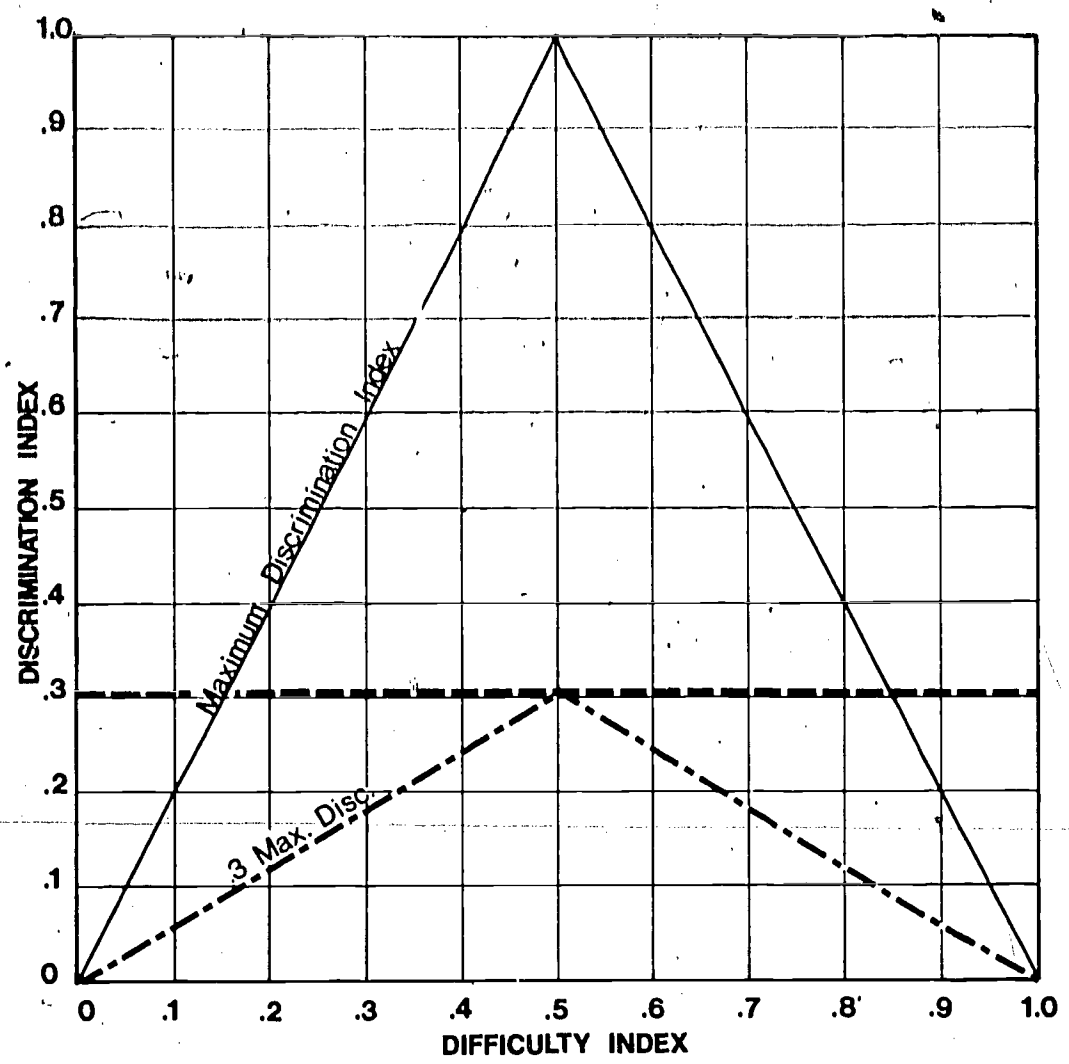
A widely used criterion for discrimination is an index greater than +0.30. Others suggest using a cutoff value of +0.20 or +0.40. Clearly the index must be positive or an item is not contributing to the central purpose of the test. Items with negative discrimination indices must be inspected very closely and either modified for future use or eliminated. Items above the arbitrary cutoff (e.g., +0.30) are considered to be adequately discriminating items, while those with values between 0 and +0.30 are considered weak discriminators. A very small proportion of items in most tests will have discrimination indices above +0.60. Instead of instituting a rigid cutoff for the item discrimination index, teachers should choose a value to begin with and modify it as necessary.

Although the item discrimination index has been treated as a measure of the item as a whole, it is really associated with only the "correct" option. While separate discrimination indices are not normally calculated for each option, considerable additional information about an item's overall performance can be obtained by inspecting the relative proportion of High and Low students choosing each of the distracting options. More Low students than High should choose any of the distractors, and any distractor that attracts a greater number of students from the High group than from the Low group should be closely examined and revised.

Item discrimination is strongly affected by the difficulty of the item. Using halves to form the High and Low groups, the following data were calculated with a sample of 100 students. The maximum item discrimination (MID) was calculated with the widest possible difference between the H and L values to produce the greatest degree of discrimination. In the chart are included the H, L and MID values for item difficulty values at every 10 percentage points between 0 and 100%.

| Difficulty | 00 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 00 | 10 | 20 | 30 | 40 | 50 | 50 | 50 | 50 | 50 | 50 |
| L | 00 | 00 | 00 | 00 | 00 | 00 | 10 | 20 | 30 | 40 | 50 |
| MID | 00 | .20 | .40 | .60 | .80 | 1.00 | .80 | .60 | .40 | .20 | 00 |

The MID peaks at the difficulty level of 50% and decreases linearly on either side to 0 values at both extremes--00% and 100%. This relationship is shown graphically in the figure following. Using a cutoff value of +0.30 for item discrimination automatically eliminates items with difficulty indices above .85 and those below .15.




This cutoff value must be sensitive to the variation of maximum item discrimination with item difficulty values. A cutoff value of .30 MID reflects such sensitivity. For instance, an item with a difficulty level of 90% would have to satisfy a cutoff value of .30 X .20, or .06. Although this complicates the use of the item discrimination index, it does provide for a criterion independent of the difficulty level of the item. The MID values for each and every difficulty level can be determined either by interpolating between the values given on the chart presented before or, more directly, by using the formula on which the MID is based.

In addition to the empirical analysis of an item in terms of the difficulty and discrimination indices, the teacher can gain considerable insight from a detailed analysis of the entire item. Each of the options should be plausible and possible, and the options keyed "correct" must be undeniably the best of the options and indisputably accurate in every sense.

## Test Analysis

Many techniques have been suggested for analyzing and evaluating tests. Three constructs commonly used to evaluate test instruments are validity, reliability and useability.

### Content Validity

The validity of a test is commonly defined as the degree to which a test measures what it is designed to measure within a given population. Content validity is based on what qualified professionals can determine by examining the test itself, its table of specifications and method of development. Generally, no statistics are involved with statements about content validity unless a percent of agreement among experts' opinions is calculated.

By examining a test and its table of specifications or course outline, the relevance, balance and specificity of the examination may be determined. These three qualities, which are part of the content validity, are defined by Payne (72) as follows:

| | |
|---|---|
| Relevance | Relevance is the quality of an educational achievement test that relates the behavior required to respond correctly to a test item and the purpose or objective in writing the item. The test item should be directly related to the course objectives and actual instruction. When used in conjunction with educational measurement, relevance must be considered as the major contributor to validity. |
| Balance. | The balance quality of a test is indicated by the degree to which the proportion of items testing particular outcomes corresponds to the "ideal" test. The framework of the test is outlined by a table of specifications. |
| Specificity | If subject matter experts should receive perfect scores (objectivity) then test-wise but course-naive students should receive near chance scores, thus indicating course-specific learnings are being measured. |

If a test is deemed to be relevant, balanced, and specific to its expressed purpose and population, it can be described as having content validity. Every test should be scrutinized by its developer(s) and qualified colleagues to insure that the test has clearly established content validity.

Statistical Validity

Criterion-related validity includes all attempts to compare results from the test in question to results of other tests designed to measure the same objectives. The forms of criterion-related validity include concurrent validity--when the tests are administered at the same time, and predictive validity--when the test in question is compared with some future test performance.

If scores from two measures for a sample of students are available, a correlation coefficient can be calculated by hand or by any one of a large number of calculator or computer programs. Few guidelines exist for interpreting the "goodness" of correlation coefficients, but the following scale is useful:

| Coefficient | Interpretation |
|---|---|
| 00--------+.20 | Indifferent, Negligible Relationship |
| +.20--------+.40 | Low Correlation, Present, But Slight |
| +.40--------+.70 | Substantial, Marked Relationship |
| +.70--------+1.00 | High, Significant Relationship |

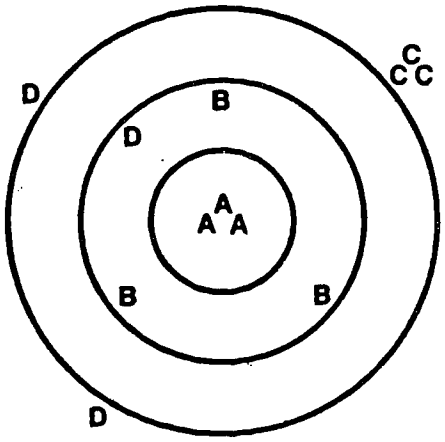
For a different mode of evaluation, the correlation coefficient may be squared, yielding a value which then represents the amount of variance common to the two measures. For instance, a coefficient of +0.50 indicates that 25% (.50 x .50) of the variance of one measure is accounted for by the other measure.

Construct Validity

The third form of validity--construct validity--assesses the degree to which some related trait or quality (construct) is reflected in the performance on the test in question. This form is used when there is no criterion measure available. Based on past research and related theory, a variable is selected which can be hypothetically related to student performance on the test being developed. Some relationships that could be hypothesized include IQ and reading achievement, or psychomotor ability and performance on a science laboratory exam. The majority of these relationships can be quantified by means of some kind of correlation coefficient, but more sophisticated statistics may be appropriate for some relationships between constructs and test performance.

Reliability

The reliability of a test is an indication of how consistently a test measures what is measured. This is also called "precision" of measurement, to differentiate it from the "accuracy" analogue for validity. An example from target shooting can be used to compare and contrast the ideas of validity and reliability.




In "test jargon," the marks in Group A are both valid and reliable; they are where they are supposed to be (the middle), and they are close together in a tight group. Group D is neither valid nor reliable since the marks are off center and widely separated. Group B is valid but unreliable, since the marks are around the bull's-eye but are widely separated. Conversely, Group C is described as invalid but highly reliable in that the marks are way off center although tightly grouped. From this analogy, it is clear that test validity is more important than reliability, though both are valuable characteristics of good tests.

Reliability is described simply as a consistency of measurement, but this consistency can be across time (called stability), in terms of form (called equivalency), or within one administration of one test (called internal consistency). Reliability across time is usually computed from a test/retest administration of a given instrument, often with two weeks or less between the two administrations. A greater delay poses questions of transfer and retention. Reliability in terms of form is accomplished by administering two parallel forms of a given measure. For both stability and equivalence, correlation coefficients between the sets of scores for a given sample are calculated for an estimate of the reliability of the test in question.

The most frequently used type of reliability--internal consistency--can be estimated from a variety of formulas and techniques. Most test analysis programs routinely calculate an estimate of test reliability using the Kuder-Richardson Formula 20. The decision to use one formula or another is often based on what kinds of data are required to make the calculation, but regardless which formula is chosen, each one considered has its advantages and disadvantages. A good source of formulas for reliability--as well as for some of the other statistics mentioned here (e.g., MID indices, correlation coefficients, etc.)--is Tate's (91) Statistics in Education and Psychology: A First Course.

Criteria for reliability vary with the author and purpose of the test, but the following chart summarizes some widely accepted guidelines:

| | |
|---|---|
| .95--.99 | Very High, Rarely Found |
| .90--.95 | High, Sufficient for Measurement of Individuals |
| .80--.90 | Fairly High, Possible for Measurement of Individuals |
| .70--.80 | Okay, Sufficient for Group Measurement, Not Individuals |
| Below .70 | Low, Useful Only for Group Averages or Surveys |

Reliability coefficients do not imply any "percentage of accuracy." Although reliability coefficients receive wide attention, they are only one measure of a test's value and are influenced by a variety of factors, such as:

- Length of Test
- Discrimination of Items
- Difficulty of Items
- Range of Ability of Group

A longer test composed of items of equal quality will have a higher reliability coefficient than will a shorter test. If each item individually contributes a unit of discrimination, the greater the number of items, the greater the discrimination and therefore, the greater the reliability. Similarly, if the discrimination of the items constituting a test is increased, the reliability of the test is enhanced. Since items have maximum discrimination at the 50% difficulty level, the difficulty of items in a test will influence a test's reliability, which decreases as the average item difficulty deviates from the 50% range.

Useability

This criterion is often described as consisting of ease of administration, scoring, and interpretation. According to Payne (72), four of the ten qualities essential to a good test are speededness, efficiency, objectivity, and fairness. These four qualities, which all relate to a test's useability, are defined by Payne as follows:

| | |
|---|---|
| Speededness | To what degree are the scores on the test influenced by speed of response? For achievement tests, speed generally should not be allowed to play a significant role in determining a score, and sufficient time should generally be allowed for all or at least most examinees to finish the test. |
| Efficiency | Efficiency is here defined in terms of the number of responses per unit of time. Some compromise between available time for testing, scoring, and relevance must be made. |
| Objectivity | For a test question to be considered objective, experts must agree on the "right" or "best" answer. Objectivity then is a characteristic of the scoring of the test, and not the form (e.g., multiple-choice, true-false) of the questions. |
| Fairness | To insure fairness an instructor should construct and administer the test in a way which will allow each student an equal chance to demonstrate his knowledge. |

Unless a timed standardized test is being used, the influence of student speed of response should be minimized by providing sufficient time for most students to complete the entire test. The concern with test administration may seem mundane, but class time is a very precious commodity at all levels of schooling. Clear, simple directions and well constructed answer keys are necessary if the item format or testing procedure is new or unique. Students should be informed simply and quietly of the time remaining until the end of the period or the test time limit. Ease of scoring is important, especially when testing five or six classes with thirty students per class. Providing separate answer sheets or locating answer blanks in one of the margins of the test paper will aid in scoring responses to objective test items. Prior to administering "problem solving" items, acceptable steps and procedures must be outlined and allowances for partial credit, if applicable, should be explained. For essay questions, a model answer should be developed or a list of points made of ideas

acceptable for inclusion in the answer. These procedures will do much to aid a teacher in being more efficient, more objective and more impartial.

## Descriptive Statistics for the Science Teacher

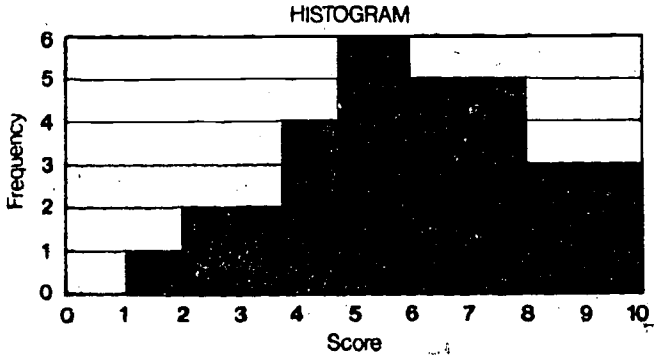
A departure point for most test analysis procedures is the formulation of a frequency distribution of test scores as shown below. On a chart listing possible scores from low to high, a tally or check mark is made whenever a given score is obtained by a student. The number of tallies becomes the frequency of that particular score.

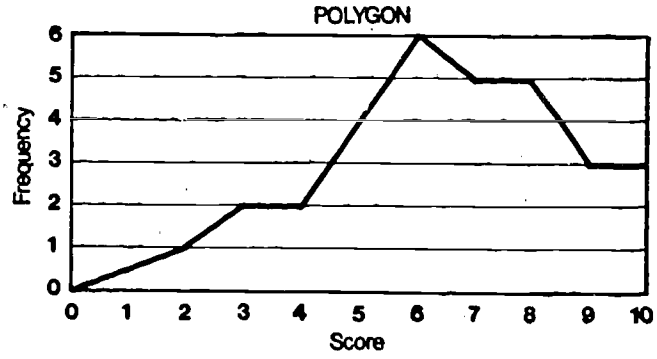
Frequency Distribution of Scores

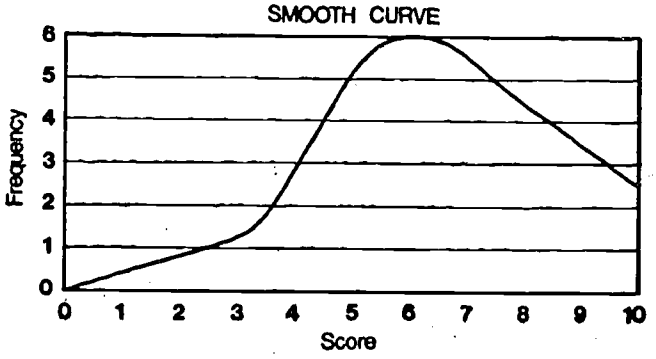
| Score | Tally | Frequency | Frequency X Score | Cumulative Frequency | Cumulative Percentage | Percentile Rank |
|---|---|---|---|---|---|---|
| 10 | 111 | 3 | 30 | 31 | 100 | 95 |
| 9 | 111 | 3 | 27 | 28 | 90 | 86 |
| 8 | 11111 | 5 | 40 | 25 | 81 | 73 |
| 7 | 11111 | 5 | 35 | 20 | 65 | 57 |
| 6 | 111111 | 6 | 36 | 15 | 48 | 39 |
| 5 | 1111 | 4 | 20 | 9 | 30 | 23 |
| 4 | 11 | 2 | 8 | 5 | 16 | 13 |
| 3 | 11 | 2 | 6 | 3 | 10 | 7 |
| 2 | 1 | 1 | 2 | 1 | 3 | 2 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |

If the deviation from the normal or bell-shaped curve is dramatic, factors that might create such a distribution should be identified and interpreted.

The central tendency and variability are often shown by means of graphical techniques such as histograms, polygons, or curves (smoothed polygons), examples of which follow.






In each of these graphic representations, the central tendency of the scores is in the 6-7-8 range, and the mode (the score obtained by the largest number of students) of the distribution is 6. If several contiguous scores (e.g., 6 and 7) are attained by the same number of students, a multiple mode can be reported. If the multi-modal scores are not contiguous, the distribution is called bimodal and the two separate scores can be listed.

From the frequency distribution can be determined the median (the score which was attained or surpassed by half of the students) or, more simply, the middle score. In the data presented earlier, the middle score (#16 of 31 students) is a score of 7, which is then called the median of this distribution. If an even number of students is involved (e.g., 30), the median is halfway between two middle scores of differing values. If the two middle scores have the same value, that value is the median.

The most frequently used measure of central tendency is the mean, which is the arithmetic average of all the scores (Mean=Sum of scores divided by the total number of students.) With the 31 scores used in the example, the sum of scores is 204, so the mean is 204/31 = 6.58. In this sample, as in most collections of scores, the mean (6.58), median (7), and mode (6) are not exactly the same, but are quite similar.

In addition to the central tendency and distribution of scores, many people are concerned with the relative position or rank of individual scores. One common parameter obtained from frequency distributions is the percentile rank of each score (the percent of scores in a particular distribution that falls below that score). Using the previous data, the percentile rank may be calculated by adding the percent of scores below each interval to half the percent of students receiving that score. For example, the percentile rank of the score "6" is calculated by adding to 30 (the percent of students with scores below 6) half the percentage of students with a test score of 6, (48 - 30/2 = 9), resulting in a percentile rank of 39. In distributions with relatively few persons, the highest score may not have a percentile rank of 100, an occurrence which is intuitively confusing. As the number of subjects increases and as relatively fewer persons achieve the top score, the maximum percentile rank approaches 100. The percentile rank, however, will always be lower than the "Cumulative Percentage" of each score due to the compounding effect of cumulation.

Many of the parameters outlined here are directly related to the determination of student grades. Grades reflect, in part, test performance, and tests are comprised of individual items. Given the impact of the net result upon our students, it is important for us to monitor item and test effectiveness even as we assess student achievement. Most of the methods for analyzing items and tests presented here may be applied easily and rapidly, and both teachers and students will benefit from their application.

# CHAPTER SIX

# Grading Students in Science

## Introduction

There are few things teachers do that are more important and visible to students and parents than the issuing of grades and evaluations. According to Link, (57) "teachers spend much time recording marks in little black books, marks which are later translated to a percent of something or a 'letter' which is a composite of something." Robinson (77) points out that:

> Although everyone suspects the reliability of grades and evaluation at one time or another, it is commonplace to hear youngsters described as 'A' students or 'C' students--as though these statements carried the same degree of certainty and 'truth' as descriptions of youngsters as brown-eyed and freckled.

These numbers, letters and written comments can make an indelible impact on a student's future achievement, interest in school, attitude toward education and life, self-concept, and appreciation of science.

Bridgham (15) collected data that supported the contention that teachers' grading practices affect enrollment in science courses. When compared with grades in other academic courses, science grades were found by Bridgham to be generally lower, with the grades of female students reflecting a greater disparity than those of male students. Bridgham concluded that if science is justified in the curriculum as being basic to a

complete general education, then student selection of science courses must not be discouraged by overly stringent grading practices. An objective, fair, impartial and accurate determination of grades is, of course, the essential challenge of all teachers.

As long as human beings are evaluating other human beings on instruments developed by human beings, "total objectivity" is impossible. A teacher can, however, strive to be as fair and impartial as possible, providing students with maximum opportunity to demonstrate their achievements. No single grading system will be appropriate for all students, all teachers, all schools, all content areas, all grade levels, all the time. A variety of systems may be combined or interchanged by flexible and imaginative teachers who assume as their goal the dynamic assessment of diversified student achievement.

Some of the most commonly employed grading systems are described as being "norm-referenced." In these systems, each student's grade is determined by how his/her achievement compares to the performance of some "norming" group. For most classroom teacher-made tests, the norm group is the class or classes whose scores are used to set the standard for grading. With curriculum project or standardized tests, the norm group is that group of students selected to validate the test. This group is often randomly selected within categories based on the size of the school and city, and often takes into account other demographic characteristics as well. This group is used to define average achievement, below average achievement, and excellent achievement in terms of percentile ranks and other scores. According to Robinson (77):

> ...a serious deficiency of norm-referenced testing is that no matter how difficult or easy the items and tests are for any group tested, there are always 'winners' and 'losers.' If excellence is defined as the upper ten or five percent of the normal curve, then 90 to 95 percent are denied excellence, and there is no way they can achieve it.

In contrast to norm-referenced systems, "criterion-referenced" grading systems are based on a performance standard describing the level of achievement with specific instructional objectives. An 80% achievement level on items written to match specific objectives is commonly expected in order to say a student has "mastered" a given unit of content. The essential difference between these two systems is the frame of reference on which the evaluation is based--in one, the performance of a group of students

determines the "norm"; in the other, a predetermined level of achievement of the instructional objectives is the "criterion" for student evaluation. The minimum level of achievement--the criterion--must be established prior to test administration or course conduct. Robinson (77) comments that "Criterion-referenced testing procedures are severely limited in the establishment of the criterion; all procedures that I reviewed were arbitrary." He does, however, point up one of the saving features of criterion-referenced systems by way of contrasting them with their norm referenced counterparts. Criterion-referenced testing procedures are, he says,

> ... intended to measure what, not how much a student has learned. ... 'Student A mastered objectives 1, 2 . . .n' and '70 percent of the class mastered five of seven objectives for the chapter' are reports of criterion-referenced tests. Such claims carry different connotations than those which proclaim that 'Bob's score on the test was 80 percent' or 'the class mean was 50 percent.'

**Absolute Standards**

Grading systems based on absolute standards are typified by the a priori establishment of some fixed criterion or distribution of specific grades. An example of an absolute grading system is one in which instructors assign letter grades to predetermined fractions of the class, as detailed in this breakdown correlating the following grades and percentages:

| | |
|---|---|
| A | Top 10% |
| B | Next 20% |
| C | Middle 40% |
| D | Next 20% |
| F | Bottom 10% |

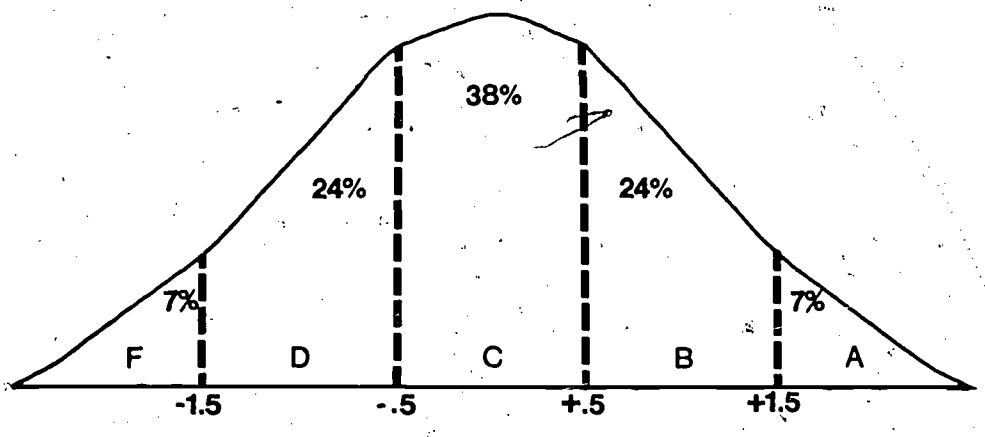
This system assumes that the achievement of all groups of students will always form a "normal" or "bell-shaped" distribution, a configuration that has long been affirmed as the ideal. Such distributions occur frequently in nature and with many human characteristics like height and weight. Other human characteristics--or rather, complexes of characteristics (as is achievement)--exhibit themselves in much less regularly plotted patterns because of the multiplicity of forces and factors shaping them. "Absolute standard" grading systems are inflexible to the possibility that an entire class could do well on any given test, an especially likely occurrence in, for instance, honors or advanced placement classes. Moreover, the imposition of absolute standards is in direct conflict with many emergent educational ideas, such as "mastery learning."

Another very popular grading system is one in which grades are assigned on the basis of a fixed proportion of the maximum possible points on a test or combination of tests. The following distribution of grades among percentages of the maximum possible scores attained illustrates one such system.

| | | |
|---|---|---|
| A | 90% or more | |
| B | between 80% and 89% | |
| C | between 70% and 79% | (Inclusive) |
| D | between 60% and 69% | |
| F | below 60% | |

These numbers may vary with individual schools and teachers, but systems like this have been passed along between generations of teachers, principals and schools like "clay tablets." Some teachers and administrators treat this kind of system as an almost sacred, inviolable law of education. Its actual origins are unclear, but its rationale is based on its simple, fair, apparently logical appearance. The limitations related to the previously mentioned system apply to this example as well.

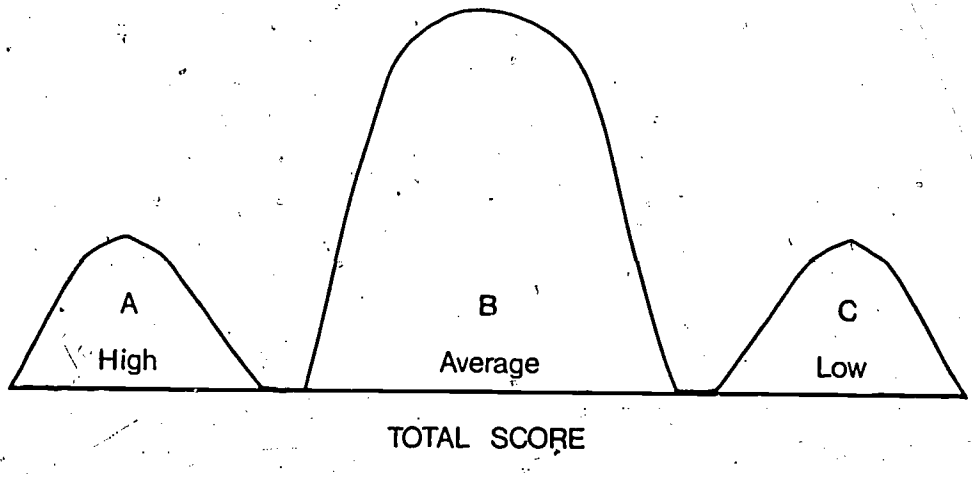
The next example--called the "normal curve" method--assigns grades depending on the number of standard deviation units a student's score is above or below the mean. This system is usually used with a final exam, but could be applied to a composite of test scores. The following figure shows one such system:




The above figure suggests that the grade of A be awarded to those having scores above the +1.5 standard deviation score, with B's between +0.5 and +1.5 SD units, etc. The percentages on the figure apply only if the distribution of scores is approximately normal. This system is described as "absolute" because of its fixed cutoff points.

**Relative Standards**

Another system found useful by some teachers is what is called the "inspection method" or, less formally, the "eyeballing" technique. This is often applied to scores from several tests and quizzes compiled over the course of an entire marking period, semester or year. The frequency distribution of these scores is examined to find gaps or breaks between clusters of scores. These gaps or breaks, if found, are used as the cutoffs between various grades. As the size of these gaps becomes larger and larger, the case becomes more convincing that some real differences exist between two groups of scores, representing groups of students. As the magnitude of the maximum possible cumulative score increases, so does the chance of finding the gaps. The location of gaps is not in itself sufficient to determine the distribution of grades, but must be coordinated with the teacher's subjective knowledge of the relative value of various degrees and kinds of achievement. In physics classes, for which many teachers normally use only A, B and C grades (because exceedingly few D and F students enroll in physics), a tri-modal distribution of scores might be most appropriate:




Many have suggested that students be graded by comparing achievement with individual learning capability. This "ability adjusted method" sounds philosophically ideal, but is fraught with many classroom problems. The major difficulty concerns selecting the measure of "learning capability" to use. Suggestions for this measure have included intelligence tests, scholastic aptitude tests, and achievement tests in the specific science area being studied. IQ tests are very widely available and provide impressive amounts of supporting statistical information, but their validity for predicting ability to achieve in a particular science course is questionable.

Achievement tests specific to the content area seem better suited to the task, but they are designed as summative evaluation devices and might measure a student's initial cognitive background poorly. Most research concludes that future achievement is best predicted by past achievement. Unfortunately, achievement in a specific content area is far easier to define and measure than is potential ability.

One way of accomplishing this task is to administer the final exam to all students prior to the beginning of the course. Then a "percentage gain score," may be calculated according to the following formula:

$$PG = \frac{\text{Post Score - Pre Score}}{\text{Maximum Possible Score - Pre Score}} \times 100$$

This system is one adaptation of the "ability adjusted method." Many of the other ability adjusted systems penalize those students who initially score high. For instance, a student who scores 85 at the outset of a course has a more difficult time improving his/her score by 10 points than does a student who initially scores 30. As a matter of fact, a phenomenon called "regression toward the mean" occurs whereby high scoring individuals, when retested, frequently score lower on the second administration--simply due to the error of measurement inherent in the test. The "percentage gain score" presents all students with a fair chance for showing improvement, regardless whether they initially score high or low.

If the final exams are reasonably valid and reliable and if the test papers are relatively secure, it is possible to consider using this kind of system. For instance, this system may be easily implemented by using the same standardized tests from year to year. Parallel forms of these tests are frequently available and would be most appropriate. In New York State, it might be logical and defensible to use last year's Regents exam as the pretest as long as no curricular changes occurred during the year.

**Multiple Standards**

Thus far, grading on cumulative total scores of similar kinds of tests has been discussed. Student grading should encompass a composite of such factors as final exams, quizzes, lab reports, lab exams, projects, papers, and others. These components can all be combined into a total score by some kind of weighting system. Students should know what this system is so that they can allocate time and energy accordingly. One hypothetical system might be the following:

| | |
|---|---|
| Final Exam | 20% |
| Semester Exam | 10% |
| Quizzes | 30% |
| Lab Exams | 10% |
| Lab Reports | 10% |
| Projects/Papers | 10% |
| Homework | 10% |

In this illustrative system, it is apparent that the values held by a teacher or school system can be conveyed through the relative emphasis given various class activities. In the above example, student performance on quizzes is very important, contributing to 30% of the final grade--as much as the final exam and semester exam combined. (The example should not be interpreted as a recommendation of that particular breakdown, but only as an example for discussion purposes.) Keeping the percentages or fractions relatively simple helps both the teacher and the students in understanding and working with the system. For students with learning difficulties it might be valuable to consider evidence of achievement from a restricted sample of data instead of the whole collection. But in the vast majority of cases, a common weighting system is a fair and just method of using evidence to support the distribution of grades.

As hand calculators, microcomputers, and time-sharing terminals become widely available, it may be possible to form composite scores by compiling standard scores rather than raw scores in effort to produce a more valid and representative picture of each student's achievement. The contribution to a student's total score is influenced more heavily by tests with greater variation, a factor usually measured by the standard deviation. As Payne (72) states, if the "standard deviation of the final exam was 20, and that of the mid-term 10, and the scores were simply added, the final exam would contribute twice as much to the composite due to the size of its standard deviation."

If this is perceived to be a problem, the most likely solution involves using standard scores. This procedure has the effect of equating the contribution of all individual scores. The basic standard score--the Z score--is the difference between a raw score ($x_i$) and the mean of that test ($\bar{x}$), divided by the test standard deviation:

$$Z_i = \frac{x_i - \bar{x}}{SD}$$

By the nature of the formula, it is apparent that half the students will have positive Z scores and half will have negative Z scores. To alleviate the "negative score" problem, many use an adaptation of the Z score, 10 Z + 50. This standard score, which is distributed about 50 as a mean, is sometimes called a T score. If Z scores are used for this procedure, however, they should be treated as a tool for internal calculations only--not as scores to be distributed to students. This recommendation stems from the difficulty of explaining to the class that one-half of them received negative Z scores. This is a function of the calculation--not a negative comment on their achievements, but it can be intuitively and conceptually confusing.

**Alternative Grading Systems**

Many authors have denounced [illegible]ool grades as being demeaning, subjective, degrading, unfair, destr[illegible], harmful and meaningless. Forgan (24) demonstrated the tremendo[illegible]s [illegible]abeling" effect of tests and grades, drawing the conclusion that "teachers (and probably students) don't want to be labeled." At the forefront of those leveling accusations against current grading practices are Sidney Simon of the University of Massachusetts Center for Humanistic Education, and Howard Kirschenbaum of the Adirondack Mountain Humanistic Education Center. They and their colleagues have written widely on the topic, including their provocative paperback, WAD-JA-GET? (47). In summary, their case is that grades do severe damage to a person's self-image, that grades have little correlation to success in a career, and that most grades are highly subjective and generally unreliable. Evidence has been gathered to support their case.

They make a distinction between private and public evaluation--the former being an

> ...important part of the learning process. It involves the teacher and student working together, sharing information and feedback, identifying strengths and weaknesses, and planning steps toward improved performance.

At the K-12 level, they recommend that parents' involvement in the private evaluation can be very helpful. On the other hand,

> Public evaluation is extrinsic to the learning process. It is the summary data about the student which is made available to parties outside the school and home--particularly to employers and other educational institutions.

Decisions made on the basis of this data can substantively affect the life of the student. Regardless of the system used for public evaluation, Kirschenbaum, Simon, and Napier (47) list four essential ingredients:

A. Clear statement of behavioral objectives, how these will be measured, and what levels of performance will correspond to what specific grades (if grades are used).

B. Meaningful written or oral communication by the teacher to the student, that considers the student's strengths, weaknesses and possible directions for improvement, with respect to the specific course objectives.

C. Student self-evaluation of strengths, weaknesses and directions for improvement, both with respect to the teacher's objectives and with respect to the student's own learning goals.

D. Time for the teacher and student to read each others' evaluations and engage in a dialogue based on this sharing of perceptions.

In an appendix to WAD-JA-GET?, the authors describe eight alternatives to traditional grading systems. They caution that these eight systems are not separate and independent but can be used in combination to develop a unique system appropriate to a particular need.

1. Written evaluations by teachers are frequently guided by forms or checklists provided by the school system. These usually have spaces for discussion of "strengths," "weaknesses," and "recommendations for improvement." Particular to science, a teacher could comment on a student's psychomotor skill development, analytical or problem solving ability, democratic behavior in lab groupings, laboratory proficiency, and scientific attitude. In addition, the teacher could share with students and parents another perspective on general academic skills, homework, notes, lab reports, class participation, library activity and skills, application of mathematical and writing competencies, etc. Such a system reflects the multidimensionality of its school and learning outcomes as well as the complexity and diversity of individual student talents and interests. However, this system does demand much more time and effort from teachers and creates extra work for those processing records in the school office. An evaluation report form developed by Kahle (44) illustrates several of these points and includes an element of self-evaluation which is further discussed next. This form appears on the following page.

"EVALUATION REPORT: SCIENCE" REMOVED DUE TO COPYRIGHT RESTRICTIONS

2. Self-evaluation has been described in WAD-JA-GET? as a process whereby "the student evaluates his own progress, either in writing or in a conference with the teacher." This was distinguished from self-grading, in which students determine their own grades. Self-grading can occur only with a prior step of self-evaluation, whether explicit or implicit. An attempt to establish student self-evaluation and grading in a junior high science class was described by Ballenger (8).

Teaching a new "junior high course in physical science, based on laboratory and hands-on activities with the emphasis on student alternatives to achieve the process objective," Ballenger felt the need to deviate from the traditional grading procedures. After discussions with his students, Ballenger developed a system of evaluation whereby students "defined their specific goals for a class period and then determined how much effort they had expended that day working toward their goals." This estimate of time spent, or effort, was converted into points that could then be transformed into letter grades. A maximum of 10 points was available for each class period, or about 5 minutes of work per point. On a weekly record sheet each student recorded the stated goal for each day, a statement of work accomplished toward that goal by the end of the class, and a point value for the time spent working that day. Students were not assigned homework or make up work, but had the option of doing activities at home, at noon, or after school for extra credit (on the basis of one point per five minutes of work). Ballenger and the students agreed that the following average daily point values deserved these respective grades:

| Average Daily Points | Grade |
|---|---|
| 10 | A- |
| 8 | B- |
| 6 | C- |
| 4 | D- |

Tests were given not as a means of determining grades, but only as a means of determining whether objectives were met. When students scored 90% or better on specially designed mastery-type process tests, they could proceed to the next activity. If they could not attain 90%, they attempted a similar process activity in another science area.

The students and Ballenger developed an instrument with 20 statements about grading to elicit some affective student responses. Ballenger

concluded that "some students will test trust to the limit, but with daily teacher interest and guidance in what they are going to do and what they have accomplished, most students will respond positively." The students' average grades improved during the grading periods in which this system was used, but Ballenger claimed this was not a function of an "easier" class or the lowering of "class standards." Rather, he felt that more material was covered at a higher level because students knew each day what was expected, what had been accomplished, and how all this influenced their grades. Ballenger concluded that this grading system--which was well accepted by parents and students--"seems fair, easy to understand, and does help to develop trust and responsibility."

3. "Give grades, but don't tell the students" does not seem a viable alternative in today's public school. Used for some time at a small private college, a "strong personalized advising system keeps students advised of their progress, informs them when they are in danger of failing, and gives them a clear perspective of how they stand in relation to their peers. . ." While this procedure may eliminate tension and competition, it seems tantamount to "throwing the baby out with the bathwater."

4. Contract grading systems are quite appropriate to science instruction with its several distinct dimensions of learning and demonstrating achievement. Most grading contracts in science classes include components of reading texts, carrying out laboratory activities and reports, viewing and reporting on audiovisual and library materials, and doing individual projects. The key elements of contract grading are the open, a priori specification of the work corresponding to a particular grade, and the students' selection of the grades they wish to earn.

The description of required and optional activities for each grade must be detailed enough for students to pursue independently, including page numbers of texts and articles, location of audiovisual materials, availability of unknowns, etc. With experience, a teacher will be able to include a sufficient variety of activities to appeal to the diverse interests of most students. The emphasis is usually on type and quantity of work, although some contracts include minimum scores on various tests as prerequisites for specific grades.

Grading contracts have been described as being either "nonbinding" or "binding." In "nonbinding" contracts, the student is merely informed about

the level of effort required for each grade. In "binding" contracts students receive penalties if they fail to meet the requirements of the grade for which they contracted. Similarly, students are required to do special extra work if they decide to try for a grade higher than the one for which they initially contracted.

The following contract developed by Kilburn (46) illustrates several elements of contract grading:

"A CONTRACT UNIT ON ROCKS AND MINERALS"
REMOVED DUE TO COPYRIGHT RESTRICTIONS.

First, for each grade, each student is required to do work from several different areas. In this case, the three areas are I. Reading and Writing; II. Rock and Mineral Identification Skills; and III. Activities. Within each area is a large number of optional activities from which a student can choose in addition to doing the required activities. Second, each student is expected to select a grade toward which to work. An important feature informs students of the demands and penalties involved in both under- and over-achievement of the originally contracted grade.

Although contract grading can help specify expected performances and relieve some anxiety and tension associated with grading, a dilemma between work quantity and quality has emerged. Merely accumulating many pages of "adequate" work should not be construed as making a superior contribution. Teachers need to wrestle with determining the quality of the effort, especially for the grades denoting excellent grasp of the subject.

5. A mastery approach to grading is not just a different grading system, but the logical reflection of an entirely different approach to teaching and learning. Much has been written about mastery learning since it was initiated by Carroll (18). The approach can be implemented with individual units, classes, or entire departments. This grading system is built on the foundation of instructional objectives which specify exactly what students should be able to do as a result of instruction. A set of such objectives defines the desired level of knowledge and skill for each unit of instruction. The teacher then must determine what to accept as evidence of mastery for these objectives. Very often a proficiency level of 75 or 80% of items sampling the domain of objectives is considered to indicate "mastery" of a given unit of instruction. Sometimes these criteria levels for mastery are based on the performance of past classes of similar backgrounds.

The mastery grading system discussed so far results in students being credited for mastering each unit when they have demonstrated the proficiency levels expected. Usually, students who do not initially demonstrate mastery of the unit objectives are expected to pursue alternative learning activities and later attempt to show mastery by means of a parallel exam on the same objectives. Such an instructional scheme assumes individualized rates of learning and measures achievement by the number of units mastered by a student. Since many schools don't appear to be prepared to incorporate this scheme, attempts have surfaced to adapt "mastery grading"

to conform to existing expectations. Some have instituted A, B, C, etc. "levels of proficiency" which are defined by the percent of achievement on a pool of items sampling the instructional domain. This bears a haunting similarity to the various "absolute" grading systems discussed earlier, the main difference being that the percentage cutoffs here are of correct items--not of students. So, theoretically, all students could display "A" level mastery of a given instructional unit.

A science teacher, Kenick (45), described her attempt to develop a grading system "really" based on mastery. Her work was grounded in her beliefs that any grade-reporting system should be "readily transferable into a form which can be utilized at other institutions" and that it should "assist the teacher in program planning and evaluation." Because of the large number of specific objectives for a given content unit, Kenick developed two grade report forms on the basis of "broad terminal objectives and more narrowly determined intermediate objectives." One form replaced the teacher's grade book while the second assumed a form that could be sent to parents. The teacher's book was "used to keep day-to-day records of homework, class work and test performances for each student." Many observations and comments about each objective could be collected there.

The grade report form provided for four levels of mastery: high, partial, low, and insignificant. A fifth level, "complete mastery," was discarded because it was "difficult to describe and impossible to measure." By assigning point values to each of these levels of mastery (e.g., high = 4, etc.) for each objective, letter grades could be easily derived. Kenick admitted that this arbitrary ranking scheme "was devised primarily because of the observed need of both students and parents to have the comfort of numbers and to bring the system in line with the traditional definitions of letter grades to which the school still subscribes." In general, this experimental program has been favorably received by students and parents. It was Kenick's hope that "a better system can be devised as people become more accustomed to noting progress rather than grade."

6. Pass/Fail grading is a form of criterion-referenced evaluation in which the teacher specifies passing requirements on levels of achievement and proficiency. A modification of the system uses the satisfactory/unsatisfactory labels. This system is based on the idea that learning may be enhanced by a more relaxed, less competitive learning atmosphere.

A Pass/Fail grading system was compared with a conventional grading system in a study involving eight chemistry classes from one high school. The classes were compared on two measures of chemistry achievement, the Science Classroom Activity Checklist, and the Attitude Toward Any School Subject instrument. Based on the results, Gatta (28) concluded that Pass/Fail grading was not a good solution to the grading problem, since "students graded on the Pass-Fail system showed significantly lower achievement of course objectives and poorer attitudes than students graded on a conventional grading system." Gatta inferred that students like to be rewarded for high achievement and will not achieve as well if this reward is missing. These results must be replicated in a variety of situations to determine if the findings from this particular school are generalizable.

7. A Credit/No Credit grading system is similar to Pass/Fail except that no student fails; she/he merely does not receive credit for the course. With a rationale similar to Pass/Fail, this system's significance is greatest for students hovering around the cutoff area. Rosen and Revak (78) discussed the use of this grading procedure with members of their physical science course designed for nonscience-oriented students at the University of Illinois. They felt a traditional grading scheme did not fit the kinds of objectives and students they were encountering. The students talked about a science course in terms of having "had" it, as though they had had the measles for a while, but managed to recover. Rosen and Revak found their new evaluation permitted different, more appropriate kinds of learning and inspired more positive responses.

8. "Blanket Grading" is a modification of Pass/Fail and a form of contract grading. The teacher specifies that "anyone in the class who does the required amount of work will receive the blanket grade." That grade is most frequently a B, but sometimes A or C. This system seems ill-suited to most schools' grading policies.

Grading is indeed a complicated process involving many subtle nuances and requiring many adaptations. All changes should be thoroughly investigated, tested, and implemented carefully and deliberately. Coordination with school administrators and communication with students and parents are essential to the success of any modification in grading procedures.

# SELECTED REFERENCES


1. Aiken, Lewis R., Jr. and Aiken, Dorothy R. "Recent Research on Attitudes Concerning Science," Science Education, Vol. 53, No. 4 (1969), 295-305.

2. Airasian, P. Physical Sciences Questionnaire. Unpublished test, University of Chicago, 1967.

3. Allen, Hugh, Jr. Attitudes of Certain High School Seniors Toward Science and Scientific Careers. New York: Bureau of Publications, Teachers College, Columbia University, 1959.

4. Allen, L.R. "Science Laboratory Student Progress Report: Sections A and B" (University Laboratory School, University of Hawaii). In Sund, R.B. and Picard, A.J. (eds.). Behavioral Objectives and Evaluational Measures: Science and Mathematics. Columbus, OH: Charles E. Merrill Publishing Co., 1972.
Checklist reprinted with permission.

5. Anderson, Richard C. "How to Construct Achievement Tests to Assess Comprehension," Review of Educational Research, Vol. 42, No. 2 (1972), 145-170.

6. Anderson, Ronald D. "Using the Laboratory to Teach the Nature of Science," The American Biology Teacher, Vol. 30, No. 8 (Oct. 1968), 633-636.

7. Attitude Toward School (K-12). Los Angeles: Instructional Objectives Exchange, 1972.

8. Ballenger, Carl R. "Trust as a Basis for Course Evaluation: An Alternative to Teacher Judgments," The Science Teacher, Vol. 41, No. 4 (April 1974), 33-34.

9. Barker, F. and Frederick, R. "Computerized Test Generation in Secondary Science Teaching," Science Teachers Bulletin, Vol. 41, No.2 (Fall 1976), 35-37.

10. Belt, Sidney L. "Measuring Attitudes of High School Pupils Toward Science and Scientists," Research Memorandum 59-14. Princeton: Educational Testing Service, 1959.

11. Billeh, V.Y. and Zakhariades, G.A. "The Development and Application of a Scale for Measuring Scientific Attitudes," Science Education, Vol. 59, No. 2 (1975), 155-165.

12. Birnie, Howard. "Identifying Affective Goals in Science Education," The Science Teacher, Vol. 45, No. 9 (Dec. 1978), 29-33.

13. Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., and Krathwohl, D.R. Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: David McKay Co., 1956.

14. Bloom, B.S., Hastings, J.T., and Madaus, G.F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

15. Bridgham, Robert G. "Grading and Enrollments in the Sciences," The Science Teacher, Vol. 40, No. 7 (Sept. 1973), 41-42.

16. Burmeister, Mary Alice. "The Construction and Validation of a Test to Measure Some of the Inductive Aspects of Scientific Thinking," Science Education, Vol. 37 (March 1953), 131-140.

17. Butts, David. Designs for Progress in Science Education. Washington, DC: National Science Teachers Association, 1969.

18. Carroll, J.B. "A Model of School Learning," Teachers College Record, Vol. 64 (1963), 723-733.

19. Cooley, W.W. and Klopfer, L.E. TOUS, Test on Understanding Science, Form W. Princeton: Educational Testing Service, 1961.

20. Doran, Rodney L. "Measuring the 'Processes of Science' Objectives," Science Education, Vol. 62, No. 1 (1978), 19-30.

21. Eglen, J.R. and Kempa, R.F. "Assessing Manipulative Skills in Practical Chemistry," The School Science Review, Vol. 56 (1978), 261-273. Checklist reprinted with permission.

22. Eiss, A.F. and Harbeck, M.B. Behavioral Objectives in the Affective Domain. Washington, DC: National Science Teachers Association, 1969.

Adapted and cited throughout Chapter 3 with permission.

23. Farmer, Walter A. and Farrell, Margaret A. Systematic Instruction in Science for the Middle and High School Years. Reading, MA: Addison-Wesley, 1980.

24. Forgan, Harry W. "Teachers Don't Want to be Labeled," Phi Delta Kappan, Vol. 55, No. 1 (Sept. 1973), 98.

25. Fraser, B.J. "Development of a Test of Science-Related Attitudes," Science Education, Vol. 62, No. 4 (1978), 509-515.

26. Fuhrman, M., Lunetta, V., Novick, S., and Tamir, P. Technical Report 14, The Laboratory Structure and Task Analysis Inventory (LAI): A Users' Handbook. Science Education Center, University of Iowa, August 1978.

27. Gallagher, James J. and Korth, Willard W. "Attitudes of Seniors Concerning Science," ERC Papers in Science Education, No. 6. Cleveland, OH: Educational Research Council of America, 1969.

28. Gatta, Louis A. "An Analysis of the Pass-Fail Grading System as Compared to the Conventional Grading System in High School Chemistry," Journal of Research in Science Teaching, Vol. 10, No. 1 (1973), 3-12.

29. Gephart, W.J., Ingle, R.B., and Marshall, F.J. (eds.). Evaluation in the Affective Domain. Bloomington, IN: Phi Delta Kappa, 1977.

30. Grobman, H. Developmental Curriculum Projects: Decision Points and Processes. Itasca, IL: Peacock, 1970.

31. Gronlund, Norman E. Stating Behavioral Objectives for Classroom Instruction. New York: The Macmillan Co., 1970.

32. Guidelines for Self-Assessment of Secondary School Science Programs. Washington, DC: National Science Teachers Association, 1978.

33. Guttman, L. and Schlesinger, I.M. "Systematic Construction of Distractors for Ability and Achievement Test Items," Educational and Psychological Measurement, Vol. 27 (1967), 569-580.

34. Haney, Richard. "The Development of Scientific Attitudes," The Science Teacher, Vol. 31, No. 10 (Dec. 1964), 33-35.

35. Harbeck, Mary Blatt. "Instructional Objectives in the Affective Domain," Educational Technology, Vol. 10 (Jan. 1970), 49-52.

36. Hofman, Helenmarie H. "An Assessment of Eight Year Old Children's Attitudes Towards Science," School Science and Mathematics, Vol. 77, No. 8 (Dec. 1977), 662-670.

37. Hofstein, A., Lunetta, V., and Giddings, G. "Evaluating Science Lab Activities." Paper presented to the National Science Supervisors Association Annual Meeting, NSTA National Convention, Anaheim, CA, March 1980. (To appear in revised form in The Science Teacher, January 1981.)

38. Humphreys, D.W. and Townsend, R.D. "The Effects of Teacher- and Student-Selected Activities on the Self-Image and Achievement of High School Biology Students," Science Education, Vol. 58, No. 3 (1974), 295-301.

39. Hurd, Paul deHart. "Scientific Enlightenment for an Age of Science," The Science Teacher, Vol. 37, No. 1 (Jan. 1970), 13-15.

40. Improving the Classroom Test. Albany, NY: The University of the State of New York, State Education Department, 1968. Adapted and cited throughout Chapter 2 with permission.

41. Jeffrey, J.C. "Evaluation of Science Laboratory Instruction," Science Education, Vol. 51 (1967), 186-194.

42. Jewett, A.E., Jones, L.S., Luneke, S.N., and Robinson, S.M. "Educational Change Through a Taxonomy for Writing Physical Education Objectives," Quest, Monograph XV (1971), 32-38.

43. Johnson, Donald H. CAGEE--An Approach to Computer-Assisted Generation and Evaluation of Examinations. Brockport, NY: SUNY College at Brockport, 1973.

44. Kahle, Jane Butler. Teaching Science in the Secondary School. New York: D. Van Nostrand Co., 1979.

45. Kenick, Lois. "A Grade Reporting System Really Based on Mastery," The Science Teacher, Vol. 40, No. 6 (Sept. 1973), 43-45.

46. Kilburn, Robert E. "A Contract Unit on Rocks and Minerals." In Romey, William D. Inquiry Techniques for Teaching Science. Englewood Cliffs, NJ: Prentice-Hall, 1968.

47. Kirschenbaum, H., Simon, S.B., and Napier, R.W. WAD-JA-GET?: The Grading Game in American Education. New York: Hart, 1971.

48. Klinckmann, E. "The BSCS Grid for Test Analysis," BSCS Newsletter, Vol. 19 (1963), 17-21.

49. Klopfer, Leopold E. "A Structure for the Affective Domain in Relation to Science Education," Science Education, Vol. 60, No. 3 (1976), 299-312.

50. Klopfer, L.E. "Evaluation of Learning in Science." In Bloom, Hastings, and Madaus, op. cit.

51. Korth, Willard W. Life Science Process Test, Form B. Cleveland, OH: Educational Research Council of America, 1968. All items reprinted with permission of Willard W. Korth.

52. Kozlow, M. James and Nay, Marshall A. "An Approach to Measuring Scientific Attitudes," Science Education, Vol. 60, No. 2 (1976), 147-172.

53. Krathwohl, D.R., Bloom, B.S., and Masia, B.B. Taxonomy of Educational Objectives, Handbook II: Affective Domain. New York: David McKay Co., 1964.

54. Kreitler, H. and Kreitler, S. "The Role of the Experiment in Science Education," Instructional Science, Vol. 3 (1974), 75-88.

55. Kruglak, H. "The Measurement of Laboratory Achievement," Parts I-III, American Journal of Physics, Vol. 22 (1954), 442-451, 452-463; Vol. 23 (1955), 82-87.

56. Likert, R. "A Technique for the Measurement of Attitudes," Archives of Psychology, Vol 22, No. 140 (June 1932), 1-55.

57. Link, Francis. "Toward a More Adequate System of Evaluation in Science," The Science Teacher, Vol. 34, No. 2 (Feb. 1967), 21-22.

58. Lowery, Lawrence F. "Development of an Attitude Measuring Instrument for Science Education," School Science and Mathematics, Vol. 66, No. 5 (May 1966), 494-502.

59. Lunetta, Vincent N. and Tamir, Pinchas. "Matching Lab Activities with Teaching Goals," The Science Teacher, Vol. 46, No. 5 (May 1979), 22-24.

60. Mager, Robert F. Developing Attitude Toward Learning. Palo Alto: Fearon Publishers, 1968.

61. Mager, R.F. Preparing Instructional Objectives. Palo Alto: Fearon Publishers, 1962.

62. Making the Classroom Test: A Guide for Teachers. Princeton: Educational Testing Service, 1973.

63. Marshall, J.C. and Hales, L.W. Classroom Test Construction. Reading, MA: Addison-Wesley, 1971.

64. Mayer, V.J. Unpublished Evaluation Instruments in Science Education: A Handbook. Columbus, OH: ERIC Science, Mathematics, and Environmental Education Clearinghouse, 1974.

65. Moore, M.R. "The Perceptual-Motor Domain and a Proposed Taxonomy of Perception," Audio Visual Communications Review, Vol. 18 (1970), 379-412.

66. Moore, Richard W. and Sutman, Frank X. "The Development, Field Test and Validation of an Inventory of Scientific Attitudes," Journal of Research in Science Teaching, Vol. 7, No. 2 (1970), 85-94. Reprinted by permission of John Wiley & Sons, Inc.

67. Nay, Marshall A. and Crocker, Robert K. "Science Teaching and the Affective Attributes of Scientists," Science Education, Vol. 54, No. 1 (1970), 59-67.

135

68. Nunnally, Jum C. Educational Measurement and Evaluation. New York: McGraw-Hill, 1972.

69. Okey, James R. "Diagnostic Testing Pays Off," The Science Teacher, Vol. 43, No. 1 (Jan. 1976), 27.

70. Osgood, C.E., Succi, G., and Tannenbaum, P. The Measurement of Meaning. Urbana, IL: University of Illinois Press, 1967.

71. Pancella, John R. "Cognitive Levels of Test Items in Commercial Biology Examinations." Paper presented to the National Association of Research on Science Teaching Annual Meeting, Silver Spring, MD, 1971.

72. Payne, David A. The Specification and Measurement of Learning Outcomes. Waltham, MA: Blaisdell, 1968.

73. Pearl, Richard E. "The Present State of Science Attitude Measurement: History, Theory and Availability of Measurement Instruments," School Science and Mathematics, Vol. 74, No. 5 (1974), 375-381.

74. Podrasky, Edward F. "Nonverbal Assessment of Learning," The Science Teacher, Vol. 43, No. 7 (1971), 39-41.

75. Processes of Science Test. New York: The Psychological Corporation, 1962.

76. Regents High School Examination: Earth Science--Performance Test Manual. Albany, NY: New York State Education Department, 1970. Task information reprinted with permission.

77. Robinson, James T. "A Critical Look at Grading and Evaluation Practices." In Rowe, Mary Budd (ed.). What Research Says to the Science Teacher, Vol. 2. Washington, DC: National Science Teachers Association, 1979.

78. Rosen, Sidney and Revak, Robert. "A Rationale for Pass-No Record Grading," Science Education, Vol. 57, No. 4 (1973), 405-411.

79. Rotella, Sam. Unpublished Course Project for IED 534. State University of New York at Buffalo, 1972.

80. Ruda, Paul. Unpublished Chemistry Laboratory Practical Examination. Cleveland Hill High School, Cheektowaga, NY, 1979. All items reprinted with permission.

81. Sawin, Enoch I. Evaluation and the Work of the Teacher. Belmont, CA: Wadsworth Publishing Co., 1969.

82. School Science Education for the 70s. National Science Teachers Association Position Statement. Washington, DC: NSTA, 1971.

83. Shoresman, P.B. Interests and Ideas, Form AV. Urbana, IL: Elementary School Science Project, University of Illinois, November 1965.

84. Shulman, L.S. and Tamir, P. "Research on Teaching in the Natural Sciences." In Travers, P.M.W. (ed.). Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973.

85. Simpson, E.J. "The Classification of Educational Objectives in the Psychomotor Domain." In The Psychomotor Domain--A Resource Book for Media Specialists. Washington, DC: Gryphon House, 1972.

86. Singer, R.N. "The Psychomotor Domain: General Considerations." In The Psychomotor Domain--A Resource Book for Media Specialists. Washington, DC: Gryphon House, 1972.

87. Stanley, Julian C. and Hopkins, Kenneth D. Educational and Psychological Measurement and Evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1972.

88. Talesnick, Irwin. Grade 12 Chemistry Laboratory Achievement Test. Kingston, Ontario: Queen's University, 1979.
Item reprinted with permission.

89. Tamir, Pinchas. "Practical Tests Assessment Inventory (PTAI)." Unpublished paper, Israel Science Teaching Center, Hebrew University, Jerusalem, March 1980.
PTAI excerpt reprinted with permission.

90. Tamir, P. and Glassman, S. "Laboratory Test for BSCS Students," BSCS Newsletter, No. 42 (Feb. 1971), 9-13.
Experiment Set-Up, To the Student, and Examiner's Guide (p. 83) reprinted with permission.
Questions for Problems 1, 5, and 6 (pp. 89, 92, 93) reprinted with permission.

91. Tate, Merle W. Statistics in Education and Psychology: A First Course. New York: The Macmillan Co., 1965.

92. Testing and Evaluation in the Biological Sciences. CUEBS Publication 20. Washington, DC: Commission on Undergraduate Education in the Biological Sciences, Nov. 1967.

93. Thomas, K.W. "The Merits of Continuous Assessment and Formal Examinations in Practical Work," Journal of Biological Education, Vol. 6 (1972), 314-318.

94. Tuckman, Bruce W. Measuring Educational Outcomes--Fundamentals of Testing. New York: Harcourt, Brace, and Jovanovich, 1975.

95. Tyler, Ralph W. "A Test of Skill in Using a Microscope," Educational Research Bulletin, Vol. 9 (1942), 493-496.
Checklist reprinted with permission.

96. Wall, C.N., Kruglak, H., and Trainer, L.E.H. "Laboratory Performance Tests at the University of Minnesota," American Journal of Physics, Vol. 19 (1951), 546-555.

97. Wallace, Charles. ERIE Science Process Test. Syracuse, NY: Eastern Regional Institute for Education, 1969.