# Resource Consultant Training Program
# Monograph No. 2

RCTP

## Assessment Practices for Determining Instructional Level and Learning Rate

Gerald Tindal

University of Oregon, Division of Teacher Education, Special Education Area,
Eugene, Oregon, 97403-1215

Published by
Resource Consultant Training Program
Division of Teacher Education
College of Education
University of Oregon



Tindal, Gerald
*Assessment Practices for Determining Instructional Level and Learning Rate*
Monograph No. 2

**Staff**

Gerald Tindal, Program Director
Jerry Marr, Editor
Denise Styer
Donna Jost
Clarice Skeen
Mike Rebar

**Acknowledgments**

Preparation of this document was supported in part by the U.S. Department of Education, grant numbers G008715106-89 and G008715710-89. Opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement by the Department should be inferred.

Cover Design: George Beltran

# Assessment Practices for Determining Instructional Level and Learning Rate

Gerald Tindal
University of Oregon

## Abstract

*This monograph describes procedures for placing students into instructional levels (materials) and ascertaining their learning rates. Most of the content represents a distillation of information from* Classroom-based Assessment: Evaluating Instructional Outcomes *(Tindal & Marston, 1990). After establishing a general perspective, a decision-making model is presented in three sections: what behaviors to assess, how to assess them, and how to use this information to place students into instructional levels and assess their learning rates. For ease of reading, the three sections have been divided and the issues kept somewhat separate. In reality, assessment practices must reflect issues from all three areas concurrently. In the first section, three performance outcomes are considered: basic skills, content information, and finally, procedural routines (problem-solving). Most teachers will structure instruction around all three outcomes, rarely focusing on one area only. In the second section, assessment methodology is described. After reviewing different information collection routines (with teachers interactively observing students, analyzing their permanent products, and using tests/measures), administration and scoring issues are considered. Finally, in section three, specific strategies for placing students into instructional levels are presented; then a decision-making model for ascertaining learning rates is presented using two evaluative systems. In one, based on criterion-referenced information, mastery and proficiency are used for assessment of specific skills and content; in the other, goal performance is assessed periodically using an individually-referenced system that tracks improvement over time.*

## Introduction

This monograph describes assessment procedures for teachers to use in making two decisions: (a) placing students into instructional programs and (b) monitoring their rate of learning. It is broadly conceived because it must accommodate many different types of students (ages and grades), many different types of teachers (elementary, secondary, content specialists, etc.), in varying school settings throughout the state country. The problem, therefore, is to provide something that is relevant for everyone.

Let's consider a few scenarios. In these examples, we have assumed that an appropriate assessment has been made for specific specialized programs (e.g., resource rooms for learning disabled students, enrichment plans for talented and gifted students, etc.), and now we must develop an instructional program and determine if it is working.

*Scenario 1:* An elementary school teacher has a first grade student who arrived at the beginning of the school year reading fluently and demonstrating considerable proficiency in writing and math. Although the student's basic skills are very proficient, a great many of the foundation rules in reading, writing, and math still needed to be taught (i.e. pronunciation, spelling, math computatiion, and writing rules, etc.)

*Scenario 2:* An elementary school teacher is developing a program for a small group of students in grades 3 to 5 that focuses on enriching their knowledge within a variety of interest areas. The students read specialty books and are taught a wide range of information. They later complete individual projects or activities that embellish the learning from classroom teaching and reading. A number of topics are addressed in this program: earthquakes, natural habitats of animals, cultural traditions, historical events, etc.

*Scenario 3:* A high school teacher has a 14 year-old student who excels in math and has a well-developed interest in computer programming. The student is very proficient in Pascal (a computer language) and has actually helped the principal develop a data-base routine for scheduling school events and attendance. In the same school, a music teacher has 2 students who are proficient musicians: One can play the violin with sufficient grace and style to be with a chamber orchestra and the other is an accomplished pianist.

Do these teachers need to know the same information about placing their students into an instructional program? Should they follow their growth and measure their rate of learning in the same manner? Probably not. Yet, these teachers must know how to proceed through a decision-making system so they can focus on the important information in their respective circum-

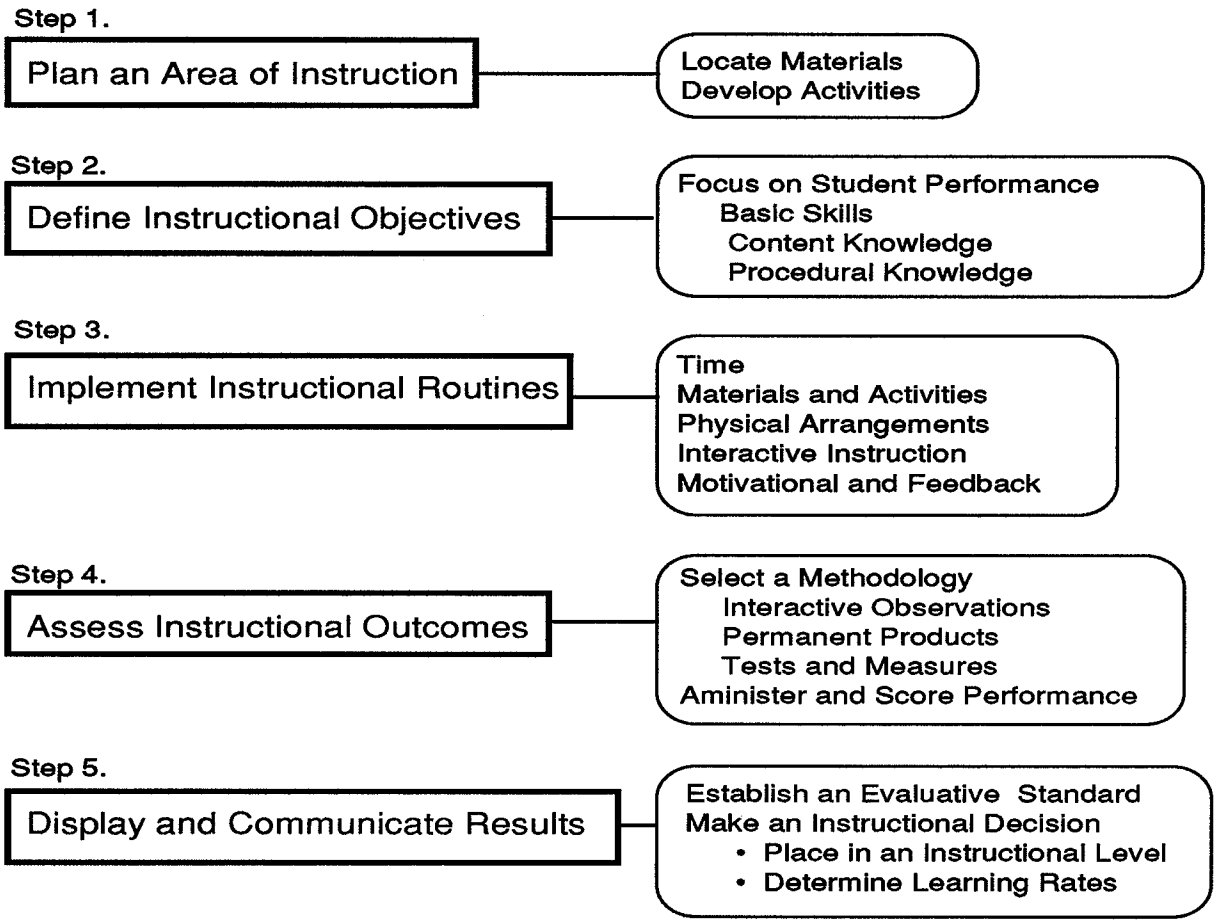


Figure 1. Outline of the Instruction-Assessment Decision-making Process

stances. Furthermore, these teachers must have some common vocabulary for communicating their programs (and the results from them) to others: teachers, parents, and specialists.

To adequately structure this task, we must frame a perspective with certain assumptions and define a vocabulary with enough clarity to provide some common ground. The assumptions will keep us from drifting into cul-de-sacs that, although interesting, are nonetheless unrelated to the task (assessing placement and rate). The language will give us the tools for making the decisions (see Key Vocabulary at end of monograph, as well as within the text).

## Assumptions

***Two decisions*** are being considered: placing students into instructional materials (which may also involve grouping students) and determining their rate of learning. Other decisions (i.e. screening and eligibility) are quite unrelated and are not considered in this context.

***Instructional level*** is comprised of the areas in which the student has adequate background and knowledge to successfully engage in material or activities, but lacks mastery or fluency in the material or activities.

***Learning rate*** implies two dimensions: (a) the material that has been learned, and (b) the amount of time needed for learning to occur.

***Classroom-based assessments*** must be used to make valid decisions about instructional placement and learning rates. Published, norm-referenced achievement tests lack (a) content validity for placement—curriculum and test items don't overlap—and (b) instructional validity—they are insensitive to instructional programs and changes in student performance.

***Teacher-derived measures of performance and learning*** are stressed in which technical adequacy (consistency and truthfulness) are ensured by careful development and implementation.

***Behavior samples*** may be diverse; however, well-developed assessments must focus on behavior, both during and after instruction.

***Quantification and qualification of student performance*** is highlighted. A wide range of information is needed, some of which is objective (requires counting amounts) and other of which is subjective (requires judgments of quality).

***Core curriculum areas*** from state departments of education are emphasized in the assessment process; however, other areas cannot be ignored.

## An Overview of the Instruction-Assessment Decision-making Process

Figure 1 depicts the five major steps in carrying out the instruction-assessment decision-making process, which are discussed below.

### *Step 1: Plan an Area of Instruction*

Materials and activities in various areas need to be selected that define the purpose of instruction. The two components of this step are selecting and adapting existing materials to fit specific instructional purposes. Typically, elementary teachers have defined areas in terms of skills (reading, math, etc.) and middle and high school teachers in terms of content. However, with integrated curricula, increasing emphasis is being given to content areas in the elementary school. Teachers from both settings, however, should be concerned with both content information and procedural knowledge.

### *Step 2: Define Instructional Objectives*

What is the domain within which teaching and learning occur? Are we teaching basic skills, content knowledge, or procedural routines. Within these areas, what is the content? For example, basic skills instruction can focus on reading, and within that domain, on several different decoding and comprehension strategies. In content areas, the focus of instruction may be on subject area specialties (history, geography, etc.) and within them specific material (the civil war, countries of the middle east, etc.). Finally, procedural routines may be math problem solving, computer programming, musical pieces, etc; and within each of these areas, instruction may focus on specific areas, respectively (specific algebraic algorithms, different languages and tasks, various compositions). Figure 2 illustrates these objectives in pyramidal form in order to hightlight their mutual interaction.

What are the critical expectations of the teacher and the relevant behaviors of the student? This step is the heart of the assessment system. To develop procedures for placing a student into an instructional program and then to measure the rate of learning, critical behaviors must be defined. You simply cannot assess what is not

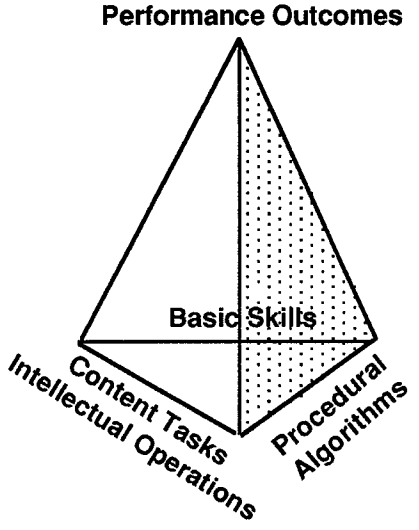


Figure 2. Three Instructional Objectives/Outcomes

visible.

### *Step 3: Implement Instructional Routines*

Although this area is implied in the discussion of assessment, I will not focus on it directly; rather it is assumed that teachers are developing and manipulating a host of strategies to deliver skills, knowledge, and routines to students. They can arrange instruction in several ways. For example, they can manipulate the time they allocate, the physical environment in which they teach, the materials they employ, the interactive teaching strategies they use, and the consequences they deliver.

### *Step 4: Assess Instructional Outcomes*

What methods should be used to assess the learner's behavior? Three different sources of information are considered:

1. Interactive observations. Teachers can collect a great amount of information about students while they are teaching by asking students questions and watching (listening to) how they perform.

2. Permanent product analysis. In many domains, students must create a product that can then be analyzed in many different ways.

3. Tests and Measures. A formal set of either paper-pencil or structured activities can be developed to assess whether students know certain information or can perform certain tasks.

What standardized assessment directions and procedures should be used? In all three of the above information sources, some procedures need to be identified for collecting and coding the information.

### *Step 5. Display and Communicate Assessment Results*

Two evaluation systems are available for determining how the student has performed and/or whether the student has improved. The first, criterion-referenced evaluations, can be used to monitor mastery, using subjective judgment of competence or proficiency. The second, individual-referenced evaluations using long range goal improvement, is based on objective measures of growth on well defined behaviors.

Two different instructional decisions can then be made using either of the evaluative standards: (a) placing a student into an instructional level and (b) determining the rate of learning.

In summary, the steps outlined above should allow both systematicity and flexibility for assessing students. The process begins with teachers defining the content and goals of instruction. Teachers can and should base assessment upon the goals of instruction and not the other way around. In the next section, you will be presented with extensive information on all three types of student behavior: Skills, knowledge, and procedural routines. In each area, a number of examples are provided for defining relevant behaviors. However, don't read the descriptions and examples as an exhaustive list; they are provided only to give you a jump start on making your own decisions. It is the decision-making process that counts. Be aware, however. We place a heavy emphasis on sampling behaviors in an organized manner.

## Focus on Student Performance

In this section, basic skills, content knowledge, and procedural knowledge. are identified as quite separate outcomes. In reality, they should be considered on a continuum, with varying degrees of emphasis on each in organizing both instruction and assessment activities.

### Basic Skills

A major function of education is to provide students with the basic skills to effectively communicate with others. These basic skills are the building blocks of all subsequent learning; they are usually completed by the end of elementary school. Although some new skills may be introduced, by design or accident, in middle and high schools, the major purpose of education is to enable students to use these basic skills to acquire content knowledge by manipulating information and use procedural knowledge to solve problems.

I will limit the basic academic skills to the following academic areas: reading, spelling, writing, and math. Within each of these areas, further limitations need to be considered to keep the focus on the minimal essential skill, rather than a more complex elaboration of it. In reading, the focus is on breaking the code; in spelling, correct sequencing of letters is addressed; in writing, production and word sequencing are considered; finally, in math, math facts and minimal computation are considered basic skills. These basic skills have five essential features:

1. They comprise the most minimal unit of meaning in communication. Behavior can't be meaningfully broken down into anything more basic.

2. Basic skills involve symbol manipulation; the emphasis is less on the content of the message than on its format and structure.

3. These skills are tool movements, useful as a means to an end; therefore they serve as the basic building blocks for more complex communication.

4. Basic skills are comprised of both content and procedural knowledge. Many facts, concepts, and principles support the symbol manipulation within any communicative act ,and the skills are organized into well-developed routines held together by rules.

5. Eventually, *automaticity* is developed with the use of basic skills. That is, skills are executed fluently and smoothly, without planning and deliberation.

These five characteristics are important in confining the focus upon the basic nature of these skills. Because of the ambiguous definitions and the ease with which these five characteristics can be applied to all communication, we need to be careful in our emphasis. Of course, eventually the difference between a basic

skill and either content or procedural knowledge becomes somewhat fuzzy, since they really represent differences of degree rather than qualitatively different phenomena. An example may help. In spelling, the following rule is used to add suffixes to root words:

As you will appreciate after reading the content of the next two sections, this rule includes both concepts and principles (content knowledge) and an algorithm (procedural knowledge that defines a series of steps to follow). Yet, it fundamentally is used to construct a minimal message with meaning: a correctly spelled word. The following concepts are embedded in the rule: suffix, root word, vowel, and consonant. The rule itself actually includes two algorithms to cover instances where the rule should be applied and those where the rule should *not* be applied and/or exceptions. Reading, spelling, writing, and math are full of many other examples. The important point is that both content information, the stuff of which content knowledge is comprised, and procedural routines, rules that organize the information into sequential steps, are used to explicate basic skills. In turn, once these basic skills are well developed, then we can engage in far more elaborate communication.

In the remainder of this section, each of the basic skills is described in the following manner. First, the central issues are described, which help organize the second part, coverage of the minimal essential features and suggestions for countable aspects of the skill. In this second section administration procedures are described, which can be embedded into an interactive observation, a permanent product, or a test/measure. Finally, specific assessment strategies are briefly considered when using this format.

### *Reading*

Probably more disagreement exists about the area of reading than any other field. The major controversy appears to be in defining comprehension, which is clouded by the confounding of the reader's background knowledge, their interests, and the manner in which the assessment is conducted. Is comprehension the same as understanding (which in turn needs to be further explicated and operationalized)? How is memory part of this definition? Rather than attempt to solve the problem, we can simply consider comprehension to be reacting to material that has been read.

The research completed at the Institute for Research on Learning Disabilities, indicates that the decoding process is a very important component of the entire reading act. In fact, most of this research, in which scores of studies have been done on many different students of varying ages in elementary school, consistently reflects a very strong relationship between decoding fluency and measures of comprehension. Students who can read proficiently tend also to be better at understanding what they read.

Before I get too far along in this issue, I should make two important points: This relationship does not imply causation (i.e. teach students to read fast and they will become better comprehenders, or visa versa). The relationship is not perfect. While it is unlikely that dysfluent readers can ever be adequate comprehenders, every teacher seems to report a student who can read fluently but cannot comprehend. Therefore, the relationship should simply be taken as a general and robust generalization, useful most of the time for most of the students, but not written in stone.

**Minimal essential features for assessing performance.** The biggest advantage to this measurement system is that it represents the terminal behavior. Unlike many of the basals, where reading has been subdivided into many subcomponents (many of which involve little reading but simply identification of word parts using a multiple choice response), this measure actually incorporates reading into the assessment process. If students can read the words correctly, then we can assume that they know the rules underlying the process. As a consequence, we can also be quite diagnostic with our assessment, listening to the prosodic features of the reading (voice quality and rhythm) and identifying the error types. The biggest disadvantage to this system is that it may have limited utility for talented and gifted students beyond the early elementary years. However, we may consider it even as an occasional check, to ensure that students throughout the elementary school years are reading proficiently and accurately.

**Assessment strategies.** The minimal essential features of reading then can be considered as decoding fluency and reactions to material. In learning to read, clearly the written message must be decoded, translated into speech (either vocal or subvocal) before any reactions can occur.

Given our limited definition of reading as a basic skill, the only measure we need to be concerned with is the decoding component. Comprehension, or reacting (as I have confined it) is addressed in the next two sections involving content and procedural knowledge. Oral reading fluency, therefore serves as the major reading outcome (silent reading is impossible to assess with any reliability). This measure is easily incorporated into the classroom and can be done quite efficiently. As conducted in most research studies, it is accomplished by the following steps:

1. A representative passage is selected; both a student passage (unnumbered) and a follow-along passage (numbered with a count of the cumulative words written after each line on the right margin) are developed.

2. Directions are established that emphasize reading carefully, not simply speed reading.

3. The administration is timed for one minute so that both rate and accuracy are considered. Remember,

basic skills emphasize automaticity, not simply accuracy.

4. Specific errors are counted as the student reads: Omissions, insertions, hesitations, and substitutions. In fact, most errors tend to be the last type; simply noting an error may be as adequate as noting its type.

5. At the end of one minute, a slash is placed on the passage where the student has read. The student is then told to read the remainder of the passage silently (to read through enough of the story to complete a retell).

6. Words read correctly and incorrectly in one minute are counted. This measure represents the basic skill of reading.

In summary, standardized administration and scoring procedures are employed to determine the student's oral reading fluency. As mentioned earlier, this skill is strongly related to comprehension. For most successful students, the measure is useful in the elementary years. Although these students may be quite high-achieving, they need to break the code just like everyone else. Denying them this skill, which is fundamental to later manipulation of information, is poor educational practice.

## *Spelling*

Probably the most controversial debate in spelling instruction and assessment focuses on the consistency of the English language. If it is consistent, then we can present instruction and assessment strategies using phonically regular words. However, if it is inconsistent, then we should not teach and assess students' skills in following limited, rule-governed spelling, but ensure that they can spell the words they are most likely to come into contact with, or high frequency words. In the first approach, which is linguistic, the stress is on phonological, morphological, and syntactical rules, with a consistent scope and sequence reflecting a number of phonological generalizations. The second approach is based on the frequency of word usage: More frequently appearing words are introduced first, followed by those used less often. Therefore, to a considerable degree, the structure of the English language dictates the spelling difficulty of individual words, with these two different factors having an influence: rule consistency and frequency of occurrence.

**Minimal essential features for assessing performance.** Given a system for sampling words, the major issue is the *scoring strategy* (we have assumed that production responses are generated). Although most norm-referenced spelling tests use selection responses, the typical classroom focus is on students actually spelling words.

Words spelled correctly, as the name implies, involves the correct spelling of an entire word. Probably the only issue with this strategy focuses on administration of homophones (i.e., blue and blew, hi and high, wait and weight). To provide an adequate reflection of student skill and not simple confusion from similar sounding words, any dictation tasks must use the targeted words within sentences to provide appropriate clues on which variant of the word is meant. This issue applies equally to the second scoring system described below.

An alternative scoring procedure is *letters-in-correct-sequence* , which focuses on successive pairs of letters that appear correctly together. Assuming that spelling is the correct concatenation of letters in sequence, this strategy focuses on the correctness of letter pairs. Following is an example.

Spell: "Handle"

Every word must have a beginning letter, which implicitly means that no other letter appears prior to the first. There is a blank space at the beginning of the word. If the word begins with an "H," place a carat (inverted V) over the blank space and the "H":

_____ H

If the next letter to follow "H" is an "A," the two letters "H A" are in the correct sequence; place a carat so it joins the "H" and the "A":

_____ H A

If the next letter is an "N," again the two letters "A N" are in correct sequence. Repeat step 2 for letters "N D":

_____ H A N

This process is repeated for each pair of letters until the entire word is scored. As in the blank space implicit in beginning each word, the word also must end in the correct letter being followed by a blank space.

The following is the correct way to score the entire word — HANDLE:

_____ H A N D L E _____

A misspelling of the word "Handle" as "Handel":

_____ H A N D E L (4 correct and 3 incorrect sequences).

As can be seen, with the correct spelling there are seven letters in correct sequence. For any word that is spelled incorrectly, there will be one more carat (letter in correct sequence) than there are letters in the word. As in this example, the word — Handle — has six letters in it. Therefore, there will be seven letters in correct sequence if the word is spelled correctly in its entirety. This is because of the half-point given for beginning and ending the word in a blank. Applying this rule helps speed up the scoring process for words that are consistently spelled correctly.

**Assessment strategies.** Spelling has few controversies; its definition appears to be clearly organized around the concatenation of letters in correct sequence. About the only major concern is with the sampling plan for selecting words. Many curricula present words as part of a phonetic family; yet the English language has many exceptions to every generalization. Therefore, some inclusion of frequently appearing words (especially those that are exceptions) is often needed to

supplement instruction and move students along.

This issue of sampling becomes particularly important as we look at student learning rates, where we want to maximize the amount learned in the shortest possible time. As presented in a later section on long range goal sampling, spelling lends itself well to a sampling plan in which students are presented a random sample of words from the entire grade level, providing both preview (words that have not yet been taught) and review (words that have been taught). When using these words at the beginning of the week, presenting them in a rolling dictation (present a word every 5 to 10 seconds for two minutes), it is possible to then identify the words that the student needs to study. At the end of the week, the entire measure is repeated (presenting the words in the same manner but in a different order). Student performance is summarized using the number of correct letter sequences. This system has several advantages: (a) only the words that need studying are addressed, (b) growth is relatively easy to see, and (c) the student can move through the curriculum at an individually paced manner.

### *Written Expression*

The major issue in written expression is whether the assessment employs a direct or indirect writing sample. With a direct measure, writers typically are presented with a stimulus prompt and directed to write a response expressing themselves in a particular manner. For example, the following prompts may be given to the writer: Describe an emotional reaction, recount an event, describe an object, explain a procedure, or defend a position. Direct assessment utilizes a specific and standardized administration format, scored in a prescribed manner, and reported according to a certain format. In contrast, indirect assessment requires no production response. Instead of writing, students select correct answers from a menu of options. In written communication, multiple choice formats are frequently used and focus on sentence structure, word usage, spelling, or punctuation/capitalization. Although few norm-referenced measures use direct writing samples, for most classroom purposes, such direct writing samples are preferred.

Both types of measures, *direct* and *indirect*, can use *objective* or *subjective* scoring systems. These four terms are often confused because many direct assessments of written expression often use subjective criteria—some form of rating scale on a dimension of quality. Likewise, most indirect writing measures employ an objective scoring format, with responses coded as correct and incorrect without reference to either a judgment or inference of quality. In general, direct assessment tends to focus on compositional skills and indirect assessments to concentrate on appropriate usage and convention. Direct assessments may be scored by objective means, however, by employing firm and consistent criteria for scoring a response as correct (i.e., correct word sequences, to be introduced later in the chapter). Likewise, an indirect assessment also may employ subjective criteria (i.e., sentence order and word usage) in which judgments are made.

**Minimal essential features for assessing performance.** Assuming a direct assessment has been chosen, the areas to consider are the prompts used in generating writing samples and the manner in which those writing samples are scored. The type of discourse or mode of expression and the specific writing topic used to generate writing must be considered as very critical influences upon the writing act itself, how it is expressed, and finally, the interpretations we can make.

To be consistent with most contemporary views—that writing is a multi-skill construct—prompts need to take into account the type of writing required of the student. Three types of writing generally have been delineated: (a) expressive or narrative, which is writer-oriented because the purpose is to express feelings, attitudes, and perceptions; (b) explanatory or expository, which is subject-oriented in that the aim in writing is to describe, explain, or present information; and (c) persuasive, which is audience-oriented, with the author taking a position on a topic and attempts to convince the audience. Narrative/descriptive writing has a purpose of presenting personal experience—the recounting of autobiographical information. It is closest to inner speech, can be viewed as self-expression, and is less discursive than other forms of writing. Expository writing contains a purpose of setting forth an idea that either informs or explains. Observations are related, analyses presented, and information conveyed in expository writing. Finally, persuasive writing is designed to convince others to adopt or endorse the writer's view A number of different types of prompts are available for generating such writing, including pictures, topic sentences, story starters, incomplete sentences, and reading passages.

Assessing written expression directly is difficult because of the lack of stimulus control in the assessment process, unlike that which occurs in other language arts areas. The correctness of the response for dimensions other than grammar or syntax cannot be determined. For example, a reading task can be controlled very easily if a passage presents skills that students are expected to demonstrate. In spelling, specific words dictated to students can be selected to assess certain skills. But teachers have little control over their students' writing, other than to specify the topic and the characteristics they desire. Thus, instructors can evaluate writing samples with their criteria in mind only with the proper scoring system.

Two scoring systems can be used: subjective or objective. All subjective scoring systems require criteria for judging student writing samples. This can be done by taking a sample and (a) matching it to another sample, (b) scoring it according to predefined quality,

or (c) scoring it for prominence of certain features. Three different systems for subjectively evaluating writing have been identified: General impression (holistic rating of compositional quality), analytic (separate judgments of various components of the composition, like organization, style, wording, etc.), and primary trait (judging the success with which the composition expresses a certain type of writing, like persuasive, narrative, or expository). Most rating scales have a low value of 1 and a high value of 4 to 7, with specific descriptions of what these values mean.

While subjective scoring systems are based on judgments of quality through rating scales, with the final score inferentially determined, objective systems are based on actual counts of specific characteristics. The most frequently used objective scoring systems include: fluency, syntactic maturity, vocabulary, content, and conventions. A sixth format incorporates more than one of the above and is referred to as a multiple-factor measure (employing several of the above counts together).

**Assessment strategies.** All compositions probably should be assessed both subjectively and objectively. If the prompts are varied and the writing samples are extremely brief, specific analytic and primary trait evaluation strategies might not be appropriate; rather, the compositions should then be subjectively evaluated holistically. Comparability across writing samples, a critical feature for using analytic or primary trait evaluation, simply is not possible with samples that vary. If the writing prompts are well standardized and the compositions of sufficient length, either analytic or primary trait evaluations can be used. All subjective judgments can employ a scale with 5 to 7 intervals, using either predefined criteria or, if many other students also write compositions, range finders can be sorted from among them. The subjective judgment can also be based on the change of compositions over time for each student. Finally, all compositions also can be scored according to three objective measures: words written, words spelled correctly, and words in correct sequence.

## *Math*

Math has been variously defined as numbers and numeration; variables and relationships, geometry (size, shape and position); measurement; probability and statistics; graphs and tables; and technology (including the use of calculators and computers). Additionally, four levels of cognitive process (Bloom et al., 1956) often have been addressed: knowledge, skills, understanding, and application. Others have defined mathematics differently. For example, a wide range of educators, including elementary, middle, and high school teachers, as well as academicians, have defined the following areas as basic skills in mathematics:

1. Elementary Computation (i.e. skills normally introduced in grades 1-6).
2. Advanced Computation (i.e. skills normally introduced in grades 5-8).
3. Applications (i.e. use of mathematics in problem-solving).
4. Estimation (i.e. giving "ball park" answers).
5. Measurement (i.e., using English & metric systems, perimeters, areas, volume).
6. Algebra (i.e. applying formulas, solving equations, simplifying expressions).
7. Understanding (i.e. describing rationale and logic for solutions/procedures).
8. Geometry (i.e. construct shapes, prove theorems)
9. Probability and Statistics (i.e. interpret charts and graphs, make predictions).
10. New Math (i.e. apply set language, read / write non-base 10 numerals).
11. Calculator use (i.e. use to solve computation problems).
12. Mathematics Appreciation (i.e. incorporate math into a larger social context).

**Minimal essential features for assessing performance.** Math, unlike the language arts, is entirely lawful, allowing us to develop a hierarchy of learning. Although most learning in the early grades focuses on sets, numbers/notations, the four basic operations and the properties that govern them (associative, distributive, etc.), later instruction addresses the social applications in the following three areas.

1. Algorithms express lawful relationships within a variety of contexts: algebra, geometry, trigonometry, calculus, etc.
2. Sentence solving expresses mathematical symbols in a problem form, which may be as simple as basic computation skills (e.g., 15 + 31 = ) or as complex as application of algorithms to solve novel problems (e.g., Given the area of a circle equals $9\pi$, what is the area of a square that just surrounds the circle).
3. Finally, problem solving places a social context around a problem which must be translated into a sentence solution (e.g., If a bag of chicken pellets can feed 18 chickens for 54 days, how long will it last if only 12 chickens need to be fed?).

This taxonomy is "generally hierarchical" in that students must first learn counting numbers before learning integers; understanding of integers and skill in their use is generally needed before work with rational numbers can begin. Likewise, understanding, knowledge, and skill must generally begin with set content first, establishing the base from which the remaining six content areas are derived. These areas are subsequently sequential, with knowledge, understanding, and skill needed in the following order: Number/notation system, operations, properties, algorithms, sentence solving, and problem solving. For example, it is very unlikely that a student could solve the area of the square problem above without knowledge of real numbers ($\pi$) or that the story problem about chickens could be solved without knowledge of rational numbers.

Furthermore, determining the area of a square simply requires application of an algorithm, while the story problem requires setting the problem up and then application of an algorithm. Both problem types require knowledge of certain operations and properties which govern them. The term "generally hierarchical" simply means that familiarity and minimal proficiency is needed with earlier content areas before such facility can expressed in later content areas. However, the exact level of proficiency is not clear, so an absolute mastery sequence may be inappropriate.

**Assessment strategies.** Assessment strategies in math can focus on very complex procedural routines. However, since we have confined basic math skills as math facts and the minimal computational routines, the only major issue is the manner for sampling problems and scoring performance. Two different systems are available for evaluating student performance on computation problems: (a) counting the number of problems completed correctly or (b) counting the number of digits in the correct place value. Most scoring systems utilize the entire problems as correct or incorrect; however, the number of digits in the correct place value may be more sensitive to student changes in performance. All four basic operations are analyzed for digit scoring, in addition to fractions and decimals.

### *Summary of Basic Skills Focus*

To acquire information (content knowledge) and develop facility with various procedural routines (solving problems), students need to have minimal skills within basic language arts and math areas. All students, regardless of classification (Chapter 1, LD, EMR, TAG, etc.), must become proficient (fluent) in manipulating symbols. High achieving students, particularly, however, may need to have curricula in these areas adjusted so that the two important decisions of placement in an instructional level and assessment of learning rates can be accomplished. For them, the curriculum may need to be adapted or adjusted to better allow them a broader range of skill exposure and more rapid coverage of material. Therefore, these areas are broadly presented to provide teachers with a focus on the issues surrounding the definition and assessment of basic skills, some minimal features that need to be part of any assessment system, and suggested strategies for conducting an assessment. Our definition of basic skills, however, is intentionally very confined. The remainder of this section addresses the two important areas in which most instruction is likely to occur with older students or those who are more high-achieving—content and procedural knowledge.

## Content Knowledge

We need to look at both the kind of information that we present to learners and the response we expect them to perform in manipulating this information and demonstrating facility. Therefore, the two components of content knowledge are the format in which it is represented (content task or type of information) and the manner in which it is transmitted by the learner (intellectual operation).

### *Content Task (Type of Information)*

A great amount of learning focuses on knowledge of specific content areas. Regardless of whether the content is divided into separate domains (e.g., history, geography, social studies, etc.) or integrated across domains, (e.g., health, home economics, and chemistry), I have organized information into discrete units to help focus our instruction and provide a learning framework for the learner. This information can be classified in a variety of ways.

Bloom and associates (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956; Bloom, Madaus, & Hastings, 1981) have divided information and learning into various "levels of knowledge" using the following continuum: knowledge, comprehension, application, analysis, synthesis, and evaluation. Generally, it is assumed that the latter forms on this continuum represent higher levels of knowledge; however, the differences between them may be very difficult to establish and they may actually be cumulatively more complex (Seddon, 1978).

Given the difficulties present in Bloom's taxonomy, another one may be more appropriate. The taxonomy presented in this monograph was developed by Miller and Williams (1973), Williams (1977), and Williams and Haladyna (1982). In this alternative taxonomy, information is organized into three different types: facts, concepts, or principles.

***Facts:*** *Associations between names, objects, events, places, etc., that use singular exemplars.* The unique feature of a fact is that the association is very narrow and not generalized across a range of events, names, places, etc. The following represent facts from history and literature that were part of a Gallup poll of 686 American college seniors reported in the *Eugene Register Guard* (Henry, 1989):

- William Shakespeare wrote *The Tempest* (answered correctly by only 42% of the students).
- Karl Marx stated, "From each according to his ability, to each according to his need" (23% thought the phrase was part of the *U.S. Constitution*).
- Mark Twain wrote *The Adventures of Huckleberry Finn* (this item was answered correctly by 95% of the students).
- Harry S. Truman was president when the Korean War began (14% thought John Kennedy was president).

By the way, the results from this survey were as follows: For the 49-question history subtest, 39% of the

college seniors failed; for the 38-question literature subtest, 68% failed. For the combined 87-question test, only 11% would have received a grade of 'A' or 'B'.

The biggest problem with tests like this, beyond the fact that the results are spread over the front page of the newspaper, is that only facts were tested. Very narrow questions were asked that represented associations between single exemplar objects, events, dates, etc. Facts are difficult to remember without an organizing scheme to relate them. Yet, they are the basic building blocks to more advanced information and are necessary, for example, in developing a key vocabulary that can be used to work with concepts and principles.

***Concepts:*** *Clusters of attributes, characteristics of names, or constructs.* They may be "thought of as a category of experience having a rule which defines the relevant category, a set of positive instances or exemplars with attributes and a name (although this latter element is sometimes missing)" (Martorella, 1972, p. 7). In this definition, rules provide the formulae for organizing the attributes of the concept; these attributes, in turn, provide the criteria for distinguishing exemplars from non-exemplars.

Our classrooms are full of concepts. The key vocabulary in a story, for example, can often provide many different concepts. In a story for young elementary students entitled, *Go, Team, Go!* the question is asked: "What's it like to be musher, or sled dog driver?" (National Geographic World, 1989). The same magazine then presents a story about the Bermuda Triangle. Both of these terms are concepts; the former is a concept of a particular type of person, and the latter is a concept of a geographic area.

Concepts really form the bedrock of a great amount of teaching and learning in all classrooms and are not limited to elementary school-age children. In a high school math class, students may be taught the following examples of the concept *polygon*: quadrilateral, rectangle, rhombus, trapezoid, square, and parallelogram. In the political science class, students may be learning about communism, socialism, and democracy, all of which are complex concepts.

Concepts form a major part of our daily vocabulary, in and out of schools. Consider the many different labels we use with our students: talented and gifted, intelligent, learning disabled, mentally retarded, etc. Many of these concepts are fairly poorly defined (the rules for specifying which attributes should be considered as exemplars and non-exemplars are often vague, containing contradictions). Many objects in our daily life are examples of concrete concepts: trees, stools (when does a stool become a chair?, automobiles (what is the difference between a car and a truck?), desks, computers, etc.

In addition to appearing in our general vocabulary, examples of concepts can appear also in the social and physical sciences. For example, after reading a short selection discussing the behavior of bacteria involved in the nitrogen cycle (from an American College Testing Program preparation book by Shapiro, 1983), students were asked the following two questions involving concepts (the correct selection is underlined):

*What is the main reason given for the importance of nitrogen?*

*A. Farmers need it for fertilizer.*
*B. It is an essential part of living things.*
*C. Bacteria are necessary to change it to nitrates.*
*D. Decomposers break nitrogen compounds down.*

Notice that the choices represent different characteristics of nitrogen (a concept), although the question sounds like a principle is implied (if nitrogen is present, then...). Below are two other general knowledge questions of science concepts from this book (Shapiro, 1983):

*An acid is any substance which:*

*A. is capable of donating a hydrogen (H+) to a reaction.*
*B. is capable of donating a hydroxyl ion (OH-) to a reaction*
*C. has a pH between 7 and 13.*
*D. can turn pink litmus a blue color.*

*In geology, an unconformity is:*

*A. A rupture in the earth along which movement has occurred.*
*B. a horizontal fold in the earth's surface.*
*C. a vertical fold in the earth's surface.*
*D. A place where young material is deposited on an older, eroded surface..*

***Principles:*** *If-then or cause-effect relationships.* Principles reflect relationships between and among different facts or concepts, more often the latter. Principles often reflect a dimension of time or space in which different concepts interact in predictable ways. The world of the classroom is full of principles, some of which focus on classroom behavior ("If you kids are noisy, we will not go out to recess.") and some of which deal with content materials. Here we are concerned only with the latter.

Science is a good source for identifying many principles. For example, after reading the same short selection on the behavior of the bacteria involved in the nitrogen cycle, students are asked the following question that involves prediction of a principle:

*What would result from the destruction of dentrifying bacteria over the whole world?*

*A. Nitrogen-fixing bacteria would eventually die.*
*B. Ammonia compounds would build up in the soil.*
*C. Atmospheric nitrogen would increase.*
*D. Soil would become depleted of nitrogen compounds.*

### *Intellectual Operations (Student Behaviors)*

While the three types of content tasks provide the grist of information that we present in our classrooms, we also need to identify specific student behaviors that we expect to change or manipulate. For example, what do we want students to do with these facts, concepts, or principles? Write reports? Give speeches? Draw pictures? And even if we do ask them to engage in these behaviors, how do we know if they are correct or their performance represents learning. The other half of ascertaining learning of "higher order thinking skills" is a depiction of the student behaviors that reflect appropriate manipulation of information.

Six different intellectual operations or behaviors are described: reiteration, summarization, illustration, prediction, evaluation, and application. Each operation represents a different level of information control and the manner in which it is manipulated. Generally, we view them as successive, with the lower level as reiteration and summarization and the higher levels as prediction, evaluation, and application. There is no reason, however, to believe that the scale of difficulty or sophistication applies beyond this simple dichotomous cut: The differences between the last four operations may be negligible; they all represent adequately complex manipulations.

***Reiteration:*** *Verbatim accounts of material that was instructed or read, which can include facts, concepts or principles.* The emphasis is on *verbatim.*

***Summarization:*** *Paraphrasing information presented in material instructed or read, which can include facts, concepts, or principles.* In contrast to verbatim recall, this intellectual operation allows individual student wording, which, if it is accurate, can reflect attainment.

***Illustration:*** *Presentation of new examples, by either depicting them and asking for the student to provide a label or description, or requesting that the student directly provide the new or unused example.* Only concepts and principles can be assessed with illustration items.

***Prediction:*** *Presentation of antecedent (preceding) information that normally leads to a consequence, and the student is asked to predict the consequence or outcome by employing a rule or principle.* Students can be asked to either provide the rule leading to the outcome or apply the rule to predict the outcome. Only concepts and principles can be assessed with prediction items.

***Evaluation:*** *Analysis of a situation by establishing or creating criteria to make judgments or decision.* Generally, principles are applied to evaluation operations; however, concepts may be considered (e.g., Is any special education label justifiable from a psychometric view?). Evaluation consists of both analysis of a problem or situation to determine factors that should be considered in making the decision, and weighting of each of these factors. It involves anticipating consequences of an act and then judging whether those consequences are acceptable according to certain criteria. Evaluation items require three basic steps: (a) select criteria; (b) operationalize criteria; and (c) make a judgement based on these criteria. The judgement needs to be supported by the criteria.

***Application:*** *Provision of an outcome and a request for students to establish the conditions needed to attain that outcome.* This operation is the reverse of prediction and is applicable primarily to principles. Again, concepts may be used if carefully considered.

### *Summary of Content Tasks and Intellectual Operations*

Together, the content tasks (facts, concepts, and principles) are identified and then configured so that students can respond appropriately. Remember that a *fact* is basic—it cannot be further reduced to a simpler form; a *concept* reflects attributions or characteristics that are in common; and a *principle* presents an if-then or cause-effect relationship. Often, the wording in our question or request gives a clue about the intellectual operation in which we are interested. Each task or operation comes with its own wordage for conveying the category. For example, to ensure that the question focuses on *reiteration,* it is important that the student be directed to "restate exactly...," "repeat the the statement...," etc. In contrast, when the question requires *summarization,* that is, not verbatim from the text or material, the directive may be to "summarize the results...," "review the categories...," "present the positions...," "describe or explain the material...", etc. *Illustration,* uses such phrases as "give an example...," "illustrate the point...," etc. *Prediction* employs directives like "predict the outcome...," anticipate the results...," "forecast the consequences...", etc. *Evaluation* includes terms such as "consider the criteria...," "evaluate the positions...," "interpret the criteria needed...," etc. Finally, *application* may employ phrases like "establish the conditions...", "vindicate the perspective..." "justify the outcome..." etc.

## Procedural Knowledge

In contrast to content knowledge, in which the focus is on what students know, *procedural knowledge* focuses on what students can do or know how to do (Gagne, 1985). Our definition of procedural knowledge is most like Gagne's learning hierarchies and cognitive strategies, in which multi-step problems have prerequisite skills and reflect self-monitoring in reaching solutions. At the heart of procedural knowledge are rules, established relationships that organize concepts and principles with other concepts and principles. Unlike content knowledge, in which the focus was on manipulation of information in a mono-operational situation, procedural knowledge is based on a sequence of steps, concatenated in either a linear or a branching relationship. Rules provide the glue for interrelating these steps. Some examples will show how rules do this.

One of the best sources of procedural knowledge is mathematics. In this field most problems, beyond

simple computational ones, exemplify procedural knowledge. For instance, in a complex multiplication problem in which 5-digit numbers are multiplied together, a variety of separate skills must come into play:

1. The problem is begun by multiplying successive numbers from right to left, multiplying each single value in the multiplicand by each single digit in the multiplier.
2. For products greater than 9, only the 1's digit is placed in the step; the 10's digit is added to the next single-digit product.
3. Each digit in the multiplier sets the occasion for another new product, which is placed on a line of its own that is successively offset to the left one place.
4. After all digits in the multiplier are multiplied by all digits in the multiplicand, the numbers in each column of all steps are added.
5. To ensure that the problem is solved correctly, it is checked by either re-completing each step or using an algorithm (e.g., checking by 9s).

This problem is probably more difficult to describe than to actually complete. In fact, to read or describe the problem represents content knowledge; procedural knowledge would be exemplified by presenting a series of numbers and directing the examinee to complete them.

Obviously, other math problems represent procedural knowledge. A description of a long division problem would likely result in a similar set of steps, as would story problems. In the example below, a mono- and poly-operation problem are provided with a series of steps identified to exemplify the rule-based sequence of problem solution typical of procedural knowledge.

*In 1880, 12,601,355 silver dollars were minted in Philadelphia, 5,305,000 in New Orleans, 8,900,000 in San Francisco, and 591,000 in Carson City. How many more silver dollars were minted in Philadelphia than in New Orleans?*

- *12,601,355*
  *-5,305,000*
  *7,296,355*

---

*A box of 25 comic books sold for $2.75 and one comic book sold for $.15. How much cheaper was it to buy a box than to buy 25 single comic books?*

- *25/2.75 (price per comic book when buying them by the box or.11¢)*
- *15¢-11¢=4¢*
- *25 x 4¢=$1.00*
- *25 x 15¢=$3.75*
- *3.75-2.75=$1.00*

The two story problems have the following characteristics in common: They require identifying relevant from irrelevant information; a computation problem needs to be set up with the correct numbers and using the correct operation; the computation procedures must be correctly completed; the answer must include the appropriate units.

All the math problems, whether computation or story problems as depicted above, are similar in three respects. First, they all require several steps for solving the problem; second, specific rules govern not only the manner in which a step is completed, but also the order in which it is executed. Third, self-monitoring is often a part of problem solution.

Procedural knowledge need not be limited to mathematics, however. Indeed, many sub-specialties within the hard sciences (chemistry, physics, biology, etc.) are premised upon procedural knowledge. The effects of air pressure, gravitational pull, molecular reactions, chemical interactions, velocity, and force all incorporate procedural knowledge. Once the concepts and principles of particular sciences are known, they can be used to work with other concepts and principles to solve more complex problems. For example, on the ACT preparation test (Shapiro, 1983), a short passage discusses various aspects of molecular motion in relation to the energy held by the molecules; the following problem is then listed:

*Using the definitions for heat fusion, specific heat, and heat vaporization given in the passage, solve the following problem: How many calories of heat energy are necessary to melt 100 grams of frozen alcohol, raise its temperature to boiling, and evaporate it given the following information:*

| | |
|---|---|
| *Temperature at which alcohol freezes:* | *-114°* |
| *Temperature at which alcohol boils:* | *78°* |
| *Heat of fusion of alcohol:* | *30 cal/gm* |
| *Specific heat of alcohol:* | *0.2 cal/gm/degree C* |
| *Heat of vaporization:* | *204/cal/gm* |

*To solve this, you need to go through the following steps:*

a. *To melt alcohol at -114°C: 30 cal/gm x 100 gm = 3,000 cal.*
b. *To heat alcohol to its boiling point: 0.2 cal/gm/deg x [78° - (-114°)]= 20 cal x (78 + 114) = 3,840 cal.*
c. *To boil alcohol at 78°C: 204 cal/gm x 100 gm = 20,400 cal.*
d. *Total calories: 3,000 +3,840 + 20,400 = 27,240 calories*

In this problem, procedural knowledge is exemplified by the following characteristics:

1. Concepts and principles were arranged in several steps, forming a hierarchy, in which prerequisite knowledge was incorporated into a problem solution.
2. Rules were used to organize and sequence these concepts and principles.
3. Self-monitoring was considered in solving the problem, in which information was sorted into relevant and irrelevant and interrelated in a systematic manner.

Like the physical sciences, the social sciences also can be considered in developing procedural knowl-

edge. As in the examples above, concepts and principles are organized and interrelated into a problem solution that includes several steps; rules govern their relationships and the sequence in which they are presented. For example, many essay questions require summarization of relevant information and an interpretation of the main ideas by either explaining or predicting various aspects of the content. In such examples, the examinee is required to manipulate information, as described in the section above on content knowledge; yet, in a more complex manner, she is asked to solve a problem that has two or more steps:

1. At least two intellectual operations are included: First, a summarization task is involved; second, either a prediction or an application item is presented.

2. These intellectual operations are interrelated in a rule-governed manner. That is, the summarization of information must include a wide range of information, only some of which is logically relevant for understanding the events leading up to the main idea. The prediction or application, likewise, must utilize relevant information and be supported by the arguments presented in the earlier two sections.

3. Self-monitoring can be incorporated by reviewing the answer for logical and supported arguments.

In summary, procedural knowledge focuses on how students perform rather than what they know. It includes rule-based behavior in which information and skills are interrelated in an organized way to solve multi-step problems. These rules guide behavior and provide an algorithm for solving the problems. In the sense that a solution to the problem can be reviewed for its application of the rules, procedural knowledge can include a self-monitoring component.

## Assessment Methodologies and Procedures

Whether teaching and learning focuses on basic skills, informational knowledge (content tasks and intellectual operations), or procedural knowledge, teachers must assess the degree to which students have attained mastery or proficiency. Three different strategies can be used in measuring such attainment, in which teachers can: (a) interactively observe students, attending to how they perform; (b) analyze permanent products, the outcomes from class assignments, homework, projects, etc; or (c) test students to determine how much they know or how well they can perform.

These three strategies are not exclusive of each other; rather, they all can be done in combination, complementing each other, or done alone, serving as the primary database for ascertaining learning. The formality in application of these three strategies also can be varied. Very standardized procedures may be established and implemented, or general procedures may be followed loosely. Finally, all three strategies are not limited to any particular implementation time; they can be completed at any time during or following instruction.

## Essential Features of Three Different Methodologies

The primary difference between the three assessment methods involves the type of data which they generate. Generally interactive observations focus on process issues—how students perform, while permanent products and tests/measures are outcome-oriented. The major difference between permanent products and tests/measures is simply the degree of structure implied within the assessment process. Often, very little structure is implied with permanent products, and it proceeds from the student. With tests/measures, a considerable amount of structure is implied and it proceeds from the teacher. However, all three types of assessment methods can generate measures of learning to place students into instructional materials or levels and document their rate of learning.

### *Interactive Observations*

In every instructional episode, teachers make demands of students that can be used to gauge whether and what students have learned. In teaching reading, students can be observed reading aloud, completing worksheets, writing a comprehension retell, or describing story highlights to others. In social studies, history, or molecular biology, students can be asked to perform by speaking, writing, or completing projects and assignments. In all these conditions, students can be observed and their learning assessed.

As mentioned above, interactive observations can focus on a wide range of different student behaviors, though they typically address *how* students perform. For example, we can observe how quickly students react, how long they are engaged, how they prepare themselves to complete work, how they approach problem situations, what questions they ask, etc. Interactive observations are important because of their proximity to classroom functioning; they occur while instruction is being delivered and therefore provide very direct behavior samples. However, they also suffer major problems, often as a direct result of being collected while instruction occurs. Because teachers are teaching, observation of student performance may be inaccurate. Therefore, to be useful, interactive observations should be carefully planned and implemented.

Two important variables can be observed interactively in assessing learning: time and information exchange. These two variables simply focus on what students are doing and how well they are performing during or after instruction. Both measures are intricately related to instruction and are therefore somewhat confounded (or limited) by it.

As a measure of learning, time has two unique properties. It is an element of instruction which is manipulatable *and* it is a component of learning that reflects facility. Classrooms are full of student behaviors in which time is an important element of learning.

When we observe student academic engagement, the measure of learning may be *time-on-task*. This measure has been found to correlate strongly with achievement: Students who are engaged more of the time also achieve more on academic measures. Another measure that is based on time is proficiency or *fluency*, which incorporates the number of items or problems produced in a fixed time period. This measure has been found to be an important component of basic skills. Students who are more fluent readers, spellers, and writers, achieve higher scores on more complex measures of achievement. Assuming that accurate or correct performance is the focus of all observations, these different measures simply illustrate the ease with which students complete work.

The other class of interactive behaviors focuses on student facility in manipulating information or procedural rules. The predominant means for assessing such facility is with teacher questioning. In most instructional episodes, teachers ask questions to check for understanding. If most students can answer a question correctly, the teacher can move through the lesson, adding new or elaborating on extant information. Questions can vary on a number of dimensions: the type of information (facts, concepts, and principles), the recency of material with which they focus (material that has been learned earlier or more recently), the purpose of the question (orienting the students, ascertaining knowledge, modeling information), the reaction to the answer (correcting, affirming, calling on someone else), etc. However, regardless of their focus, breadth, or intent, questions provide information about learning. In most classroom settings, with 20 to 30 students, it is difficult for teachers to keep track of individual responses. Indeed, students typically have differential opportunities to respond. However, if teachers are working individually or in small groups, question answering may become a viable assessment method.

### *Permanent Product Analysis*

Every day, students create a number of products that can be analyzed to determine if they are placed at an appropriate instructional level and whether or not they are learning. Permanent products can include published or teacher-made worksheets, assignments, projects, work samples, videotaped presentations, etc. Just about any outcome created during or following an instructional episode can be considered a permanent product worthy of analysis.

In many respects, permanent products represent the eventual goal of instruction. As defined in this monograph, they may often be student-driven (originated and completed by the student independently). They are usually completed with few prompts from the teacher and represent generalizations or extensions from instruction. The structure for designing permanent products is completely open-ended and may be fit within any instructional units. Consider the wide range of products in the list below:

- Stage drama that has been videotaped.
- Painting, sculpture, or etching on exhibit.
- Written essay on the fall of communism in the 1990s.
- Composition expressing narrative writing style.
- Story retell describing the moral of an Aesop fable.
- Solution to a math riddle.
- Computer program for solving a complex sorting task.
- Audio-taped piano composition.

This list of products could be extended indefinitely. The important point is that instruction is often oriented toward providing students with the necessary basic skills, content information, and procedural strategies to produce something important within the natural environment (beyond the school walls and part of the larger social contingencies). These products can all be analyzed for both quantity and quality.

### *Tests and Measures*

Although the backbone of most learning assessments, tests and measures have been confused in their definition and implementation. For our purposes, a test can be defined as a systematic procedure for measuring learning with correct and incorrect performance. In this definition, the two key words are systematic and measuring. Tests can be developed by anyone: teachers, instructional assistants, publishers, etc. They can also range from formal to informal in their administration and scoring. However, they must be systematic in their development and serve to measure performance. *Systematicity* simply means that the domain from which items are developed is well explicated or defined. This qualifier distinguishes most tests from worksheets, which have very little systematic sampling plans. That is, items have been selected and formatted with little formal organization. The focus on measurement ensures that behavior is somehow quantified (even quality judgments are given ordinal ratings that extend from low to high).

Tests and measures can sample any of the three areas: basic skills, informational content (content knowledge), or procedural knowledge. The steps in developing a test are simple and include the following:

1. A domain is defined for developing items. This domain can be based on pages in a book, certain types of information (e.g. vocabulary concepts or various principles such as cause and effect), types of problems utilizing various algorithms (the distributive property in algebra—(2a+3c) x (1d+2b), the Pythagorean theorem, etc.).

2. A system is established for sampling items. Usually, representative items of a certain type are

sampled according to some proportion (e.g., half the math problems include geometry sentence problems and the other half include area story problems; all key vocabulary words are used in sentences).

3. The specific items are developed and formatted for presentation to the student. A number of arrangements are possible: Items move from easy to difficult; multiple-choice items are presented first with short answer essay questions presented later, etc.

4. A scoring key is developed for measuring performance.

## Summary of Assessment Methodologies

Three assessment methods can be used in ascertaining what students know and/or how well they perform: interactive observations, permanent product analysis, and tests/measures. These three systems are meant to be complementary and need not be considered as all or nothing. Most teachers can use all of them, with one serving as the primary method and the others serving to corroborate the findings. Such a multimethod approach is probably better in ensuring the outcomes from the assessment are valid (truthful). These three approaches may also differ in their degree of standardization, ranging from informal to formal. Since this issue, however, influences the reliability of the measures, and given the need for an assessment system to be reliable, the next topic to consider is administration and scoring procedures.

## Administration and Scoring of Student Performance

In the previous sections, the focus has been on what to assess, with a range of options presented from which you can choose. Teachers need to decide what student behaviors should be assessed, using the goal of instruction as the driving force. In this section I offer a similar perspective on administration and scoring of student performance A range of options is presented; the choice should reflect the goals of instruction.

Although you may consider the most important question to be the content of assessment, the different assessment administration and scoring options are just as critical. For example, few teachers would equate an assessment focusing on content information (i.e., evaluation or application of concepts or principles) with an assessment focussing on procedural routines in a story problem, given the same methodology (i.e., analyzing permanent products). Yet, administration and scoring issues may yield results that differ as much as those which focus on different performance outcomes. The teacher who employs a standard assessment routine using an objective count of the number of problems completed correctly and incorrectly is simply not doing the same thing as the teacher who uses a subjective rating of quality using unstandardized administration procedures. Results from these two assessments, even given similar content and methodology, are not comparable. The purpose of this section is to acquaint you with the range of options, provide a basis for deciding which to use, and establish a perspective for the last and probably most important task: making instructional decisions.

### *Administration Issues*

The major concern with administration of assessment tasks is consistency. In the previous section, with our focus on different types of behaviors to assess in three contexts (interactive observations, permanent products, and tests or measures), validity, or truthfulness was emphasized. In this section on administration, or how to assess, reliability is the major concern. Basically, you can administer any assessment procedure in a number of ways that simply vary on formality, from quite informal and unstandardized to very formal and completely standardized.

The choice should reflect the emphasis of instruction. For example, with decisions that are relatively unimportant (e.g., pairing students for a lesson so that they are approximately equal in background knowledge), a more informal procedure may be selected. However, for more important instructional decisions (e.g., placing a student into a specialized program or skipping units of basic skill instruction), more formal procedures should be selected. The reason is that the assessment process is likely to result in more reliable information with more formal administration and scoring routines. And with reliability comes validity: The outcomes are likely to more accurately reflect the student's actual performance (if we had no measurement error) on important behaviors.

The choice of assessment procedures on the informal to formal continuum actually depends, therefore, upon the degree of error with which you are willing to live. Selection of informal procedures increases the likelihood that the assessment will include more error than if more formal procedures had been selected. So, the most important question is this: Where does error come from and how can it be controlled?

Generally, error can arise from any of the important components within the assessment process: (a) the person being assessed, (b) the person collecting assessment information, (c) the manner in which the assessment is conducted, and (d) the situation in which assessment is occurring. A simple example should clarify these four sources of error.

Students come to school in varying states: sleepy, alert, hungry, grumpy, uncertain, etc. If we assess them on some instructional unit to place them with other students or into some materials on that particular day, their performance may well be a function of the way they feel. Teachers also come to schools in various conditions, the same as students. On any particular day in which these assessments are completed, the results may be differentially a function of the manner in which the information was collected. For example, if you are

tired, you may take short cuts in the administration procedures. Schools also vary from day to day. The room in which assessments are conducted may be too hot or cold, or loud on the particular day in question, which in turn influences the quality of the information that is collected. Finally, the assessment instrument itself may affect the results. For example, if an interactive observation is being collected, the view of the room may be blocked or the questions that are asked may be unclear. If a permanent product is being analyzed, the directions to the student for completing the task may be unclear. With tests/measures, the quality of the print may be poor, confusing students who take the test.

All four of these sources of error influence the results, first by making them more inconsistent, and subsequently, by making them potentially less truthful. The moral of the story is that assessment procedures should be standardized as much as possible. The advantage of standardization is twofold. First, the results are likely to be more consistent and therefore more truthful. Second, the results can be communicated and interpreted more easily to others (parents, other teachers, etc.). Without standardization, the information may be quite uninterpretable, as we will see later. The only disadvantage to standardization is that it is more work and, until its profound influence is appreciated, is likely to be viewed as unnecessary.

### *Scoring Procedures*

Numerous possibilities are available for scoring any student's performance, irrespective of all issues presented yet (e.g., the content of an assessment or the process for conducting it). The central decision focuses on maximizing the sensitivity of the scale to summarize student performance. At the broadest level, we can decide between qualitative or quantitative information. Qualitative information provides rich and elaborate information on the context of student performance; however, it is often complex to interpret and difficult to assay progress. Quantitative information, though well suited for measuring progress, is often viewed as missing an important element about the manner in which students perform. In this next section, we propose that both types of information be collected routinely.

Before proceeding, however, I need to draw an important distinction. In this monograph, I have emphasized production responses in which students actively create, construct, produce, respond, etc. during the assessment process. The information that is collected can focus on either the manner in which students respond or the product of that response. In a selection response (i.e., a multiple-choice test), all we can observe is the product of the response. This focus, however, should not be confused with the dimension of quality versus quantity. We can create an assessment process that has any combination of these different dimensions, as depicted in the Figure 3.

| | | Response Mode: Production | Response Mode: Selection |
|---|---|---|---|
| Scoring System | Objective | | |
| | Subjective | | |

Figure 3. Dimensions of Assessment Process

For example, an assessment task may use a production response with an objective scoring system (like the number of words read correctly from a basal passage or the number of letters correct in a spelling task). In contrast, the production response may be subjective scores (e.g., using a rating scale of holistic judgments with a measure of written expression quality). Of course, most published measures use selection response with objective scoring (primarily because of the ease with which they may be scored). However, it is possible to develop a selection response that is *subjectively* scored, as depicted in the multiple-multiple choice responses that use a rating scale in the comprehension measures developed by the Center for the Study of Reading. In this task, students read a passage and then answer a series of questions, rating each choice on degree of relevance or possibility, according to the story.

**Subjective scoring.** Many dimensions of student performance cannot be objectively scored or counted. For example, how would you objectively score persuasiveness of a written expression composition or the degree to which a story retell is consistent with the story itself. Although some objective counts can be made of various dimensions of these two outcomes (e.g., number of adjectives in the written composition or number of characters included in the retell), they lack an important dimension for evaluating performance: A view of the whole or undifferentiated product. Many other student creations present teachers with a similar dilemma; when they are broken down into a number of subcomponents, important dimensions of the creation are overlooked and the entire piece is not evaluated holistically. Furthermore, the reactions of others are not considered, which may actually be important components of the assessment. In many written compositions, it is the reaction itself that is the end goal of instruction. Therefore, to consider student creations holistically and to assess perceptions and reactions, a subjective evaluation system is needed.

To subjectively score student performance, five issues must be resolved. before a reliable evaluation can be made. Only then can agreement exist among individuals not only about what is being evaluated, but how it is being evaluated. Of course, without such agreement (or reliability), it is not possible to have a meaningful measure (one that is valid or truthful).

First, the dimension for evaluating performance must be defined. Generally, a concept is selected along with a number of synonymous adjectives. For example, in writing, the concept may be *clarity*, and the synonymous adjectives are *flavor, elaboration, organization, spontaneity*, etc. Each of these words is quite ill-defined until further descriptors are provided. Often, other student compositions can be used to help identify a number of characteristics thought to reference the trait or concepts in question.

Second, a scale must be constructed that differentiates gradations on some continuum of more to less. This scale can have any number of points or anchors on it. The most simple is a 2-point scale, in which a decision is made about the presence or absence of something (in which case the scale can be considered a checklist). Slightly more complex is a scale that has 3 points on it, for example, high, medium, and low. Finally, at the most complex level, several points or anchors may be delineated, ranging from 4 to 7 points. Very few scales exceed 7 points, because it becomes quite difficult to differentiate among points with any consistency. Of course, as the concept or trait must be defined using a number of descriptor statements or using different synonyms or adjectives, each anchor also must be delineated (e.g., what is high or low).

Third, the judgment process must be calibrated. Usually, with this step, several exemplary papers or products that represent the range of possible scores are evaluated and compared with at least two judges (or on two occasions). In the process of making judgments, the concepts or traits as well as the anchors are refined; by comparing judgments, subtle nuances of meaning may be adjusted and clarified. In calibrating a subjective evaluation, the comparisons are simply whether the judges agree (with each other or with themselves on different occasions). The next two steps should not be completed if such agreement is not attained; rather, the traits or anchors should be simplified or clarified.

Fourth, the judge proceeds through all the student creations and assigns a value, frequently reviewing the scale to ensure a well-calibrated judgment. Usually, this step is completed with certain guiding procedures that the judges follow. For example, the judge may be told to move from one paper to the next very quickly, spending less than one minute per paper; or the judge may be told to consider only one trait or concept at a time and proceed through all papers focussing on one trait before returning through them to focus on others.

Finally, student performance is summarized, and the distribution of scores is plotted. If the anchors have been identified ahead of time, the distribution may not be normal. If the anchors are defined using student papers, the distribution should be normal, with a few students having high or low scores and most students in the middle.

In summary, subjective evaluations help quantify student performance on dimensions that are considered as a whole (in which the parts do not equal the sum) and for those characteristics of performance that are difficult if not impossible to count separately. Subjective judgments often provide a useful supplement to the evaluation process, but should not be used without also considering an objective count of performance. The primary reason for including other more objective information is that the scale for showing growth on subjective judgments is very narrow and likely to quite insensitive.

**Objective counting.** Student productions may be objectively scored by counting any relevant feature, whether correct or simply present. While we have had a long history of counting performance as correct or incorrect (on tests/measures), we actually can consider many more aspects of performance. For example, on student essays, we can count any of the content information (as well as the intellectual operations that are exhibited). The number of facts, concepts, and principles may provide an interesting distribution that reveals proficiency in manipulating the information that was taught. Creative writing compositions can include a number of different counts: number of words correctly sequenced, words of various sorts (nouns, verbs, and adjectives), types of sentence (incomplete-fragments, simple, compound, or complex), or thought units (Hunt, 1964). All of these measures reflect an important aspect of writing. Together, they essentially define writing.

The only important requirement in establishing an objective scoring system, beyond generating and using rules to ensure scoring consistency, is to make it as sensitive as possible. For example, in scoring math problems, we have traditionally scored answers as correct or incorrect. However, we can actually score the number of digits correct and incorrect. In spelling, we also have focused primarily on whether the word is spelled correctly. Yet, we could count the number of letter sequences correct and incorrect. An important area in which this concept becomes important is math story problems, where the complexity of the problems precludes many of them appearing on an sample assessment. Since an important maxim in the measurement world is to increase the number of items in order to increase the reliability of the measure, such a strategy of counting components or steps may be critical. Therefore, rather than scoring the entire problem as correct or incorrect, we could score each problem for (a) the manner in which it is set up (the correct numbers are associated together using the appropriate operation), (b) the computation that is completed, and (c) the unit that is supplied. Even with this simple modification, we have increased the scale threefold. An important side effect from using this strategy of counting components is that our attention is directed to *how* students

perform, providing us with potentially important diagnostic information.

### *Summary: Scoring Options & Examples of Assessment Methodologies*

Figure 4 depicts both types of scoring systems with all three types of content. Both subjective and objective scoring procedures can be applied to any cell within this matrix. Three columns are depicted, reflecting the three assessment methodologies: interactive observations, permanent products, and tests/measures. Additionally, three rows are presented, reflecting the different types of student performance: basic skills, content information, and problem-solving procedures. To summarize all this information, a number of examples are listed below, identified according to both dimensions. You can determine whether the scoring system should objective or subjective.

**B. Permanent product analysis of a basic skill.** The teacher has assigned students to write in a journal and is interested in analyzing the changes that occur from the planning and composing process to the final edited draft. In addition to creating a learning portfolio, in which the compositions created in the various

| | Interactive Observations | Permanent Products | Tests/ Measures |
|---|---|---|---|
| Basic Skills | A | B | C |
| Content Information | D | E | F |
| Procedural Knowledge | G | H | I |

Figure 4. Scoring Systems and Assessment Methodologies

writing phases are stored, he has analyzed the compositions for the following metrics:

1. Subjective analysis of organization and ideas (yielding two separate scores that could range from 1 to 5).

2. Number of T-units and number of words per T-unit.

3. Number of sentence types (fragment, simple, compound, and complex).

**C. Tests/measures of spelling skill.** The teacher presents a random sample of words from the entire curriculum once each week. The students study those words they have spelled incorrectly and retake the test at the end of the week. A student's performance is analyzed for the number of correct letter sequences.

**E. Interactive observation of content information.** The teacher has placed three students from Grades 4 and 5 in an instructional unit on Egypt; she is teaching from an integrated perspective, covering concepts from geography, social studies, history, math, and science. Many different facts, concepts, and principles are being covered in each of these knowledge domains. During the presentations, discussions, and guided discovery practice exercises, she asks the students questions about the material they have read (literal comprehension that utilizes reiteration and summarization) and its implications (inferential comprehension that utilizes illustration, prediction, evaluation, and application). Each day, she prepares some questions in advance that have been designed according to the content information and the intellectual operation. During the lesson, she codes the initials of the student to whom she puts the question and circles the initials if the question is correctly answered. At the end of the week, she can count the number of attempts, as well as correct and incorrect responses the student has displayed.

Of course, this same lesson format could utilize either a permanent product analysis or a test/measure. In the former, a different metric would need to be prepared. For example, the number of facts, concepts, and principles that are correctly integrated into an essay may be counted. In addition, the essay may be subjectively evaluated using a holistic judgment of writing quality and inclusiveness. A test/measure, if employed, would likely utilize the same metric as described for the interactive observation: a count of the number attempted, correct and incorrect.

**G. Interactive observation of a procedure.** This teacher has a small group of five middle school students placed for instruction in math. The emphasis of instruction is on problem solving using algebraic equations. During the lesson, he writes a problem on the board, asking the students to set up the problem; he then calls on one of them to write the response on the board. One point is marked on the student's worksheet if s/he wrote out the correct formula. The teacher then asks the students to solve the problem. He calls on a different individual to write the answer on the board, and the students critique that answer. If they conclude it is correct, they compare their answers to it and score another point on their worksheets. If a student's answer is incorrect, s/he locates where the mistake was made and circles that part of the work. At the end of the exercise, the students count up the number of points they have earned and record them on the teacher's master sheet.

**I. Test and measure of procedural knowledge.** A science teacher has just completed a unit on the chemical impulses of the brain for three seniors in high school. She has spent the last six weeks teaching her students what chemicals are involved in brain activities and how they work to stimulate various functions of perception and cognition. As a final project she has assigned them to study the chemical reactions of various popular

street drugs and describe how they disrupt normal chemical reactions. She has required them to turn in a report that summarizes the properties of the drug, the centers of the brain that are affected by its use, and to predict the uptake, influence, and dissipation of the drug on behavioral functioning. She plans to score their performances in two ways: scoring the steps in the procedural sequence of chemical reactions that was taught and scoring holistically according to the integration of the information from the instructional unit.

### *Frequency of Administration*

The last topic to be considered in administration and scoring assessments is the frequency with which we should measure. Two rules of thumb may be considered in dealing with this issue:

1. Measure no more frequently than you expect to see change.
2. You never know how much is enough until you how much is more than enough.

Let's take each of these rules and look at some of the issues.

In measurement and assessment, more is generally better than less, at least when we are talking about acquisition of academic skills. Our instructional expectations are for students to acquire increasingly more information. Likewise, our assessments tend to get more stable and reliable when conducted more frequently or when we use more items. Obviously, however, we also have some limits. Students cannot be expected to concentrate at all times of the day for long periods of time; nor do we have time to continuously assess students. So compromises are in order.

The goal is to maximize our time so our assessments occur at the same time and with the same frequency as the samples of behavior we assume represent learning. We can then determine whether learning has occurred and move on to other topics or elaborations of the current one. We don't want to assess anymore than we have to; however, we also don't want students to be caught in "down-time" (i.e., placement in material they have already mastered), since they cannot move any faster than we pace them. The trick, then, is to strike a balance between these two opposing issues, moving students through material at an appropriate pace.

When teaching students complex problem-solving strategies that incorporate many steps, utilize many concepts, and have several rules embedded in them, learning may not occur within one instructional episode. Indeed the lesson itself may span several sessions. Some learning is simply not expected to occur rapidly, from moment to moment. It would be wasteful to measure the student each instructional session. Yet, many of the basic skills can be learned within an instructional episode. Many of our reading decoding strategies, spelling generalizations, and writing conventions can be taught and learned very quickly by most students. Therefore, in determining how often to assess students, you should consider the type of performance that is being demanded and the amount of instructional time devoted to teaching it. Measurement that occurs any more frequently than expected is most likely a waste of teacher's time.

The other side of this issue, however, is that learning cannot occur until an assessment occurs, whether it is done formally or informally. To state anything more definitive is simply a fabrication. Therefore, assessments need to be timed so they don't occur when it is too late and they simply become a summative evaluation that learning has occurred. More frequent measurement would reveal that instruction could have moved at a faster pace. This issue is really at the heart of assessing learning rates. The problem is that teachers must assume responsibility for moving students at a rate in which they can succeed. Of course they cannot succeed with anything they haven't been asked to consider. Although moving a student too fast can pose a danger, it is probably less of a problem with high achieving students than it is with special education students. Therefore, the dictum described above errs on the side of demanding too much. Anything less is a waste of the student's time. Since this topic represents the content of the last section—making instructional decisions for placing the student into appropriate instructional materials and contexts and to document rate of learning—we need not get into detail here.

In summary, the assessment methodology has a vital influence on performance outcomes. Both administration and scoring procedures must be clearly developed to generate a meaningful database. Once this database is developed, we can actually employ it to make instructional decisions, placing students into instructional materials or pacing them through a program and ascertaining their rates of learning.

Up to this point, I have presented information on what to assess and how to conduct an assessment. Yet, without some guidelines on how to use the information (to ascertain placement or rate), there can be no significant improvements in the educational programs for students. Placement in an appropriate instructional level has considerable bearing on the rate at which the student is likely to learn.

## Placement in Instructional Level

An appropriate instructional level implies that the student has enough skill and information to interact with the material and/or the teaching but is not yet proficient without support. We can turn to two areas of educational research to formalize this general definition: informal reading inventories and academic engaged time. Both concepts depict a system for ensuring that students' interactions within an instructional episode (with teachers, other students, or materials) are meaningful (successful and extend current performance).

## Informal Reading Inventories

Informal Reading Inventories (IRIs) were first developed in the middle 1940s and now comprise a strong tradition both in the schools and in the reading research literature. Although we may not formally adopt an informal reading inventory as an assessment tool, we can use the logic behind it to help us make decisions about placement. An inventory is basically a procedure in which students read materials of varying difficulty and answer questions about the material they've read. Although there are many nuances differentiating various informal reading inventories, they all center upon a basic definition and set of procedures for determining levels of performance.

The major purpose of an inventory is to place students into an instructional level and to know the levels where the student can read independently or is frustrated. For young children, another level is often considered that focuses on their ability to understand material that is read to them (called listening comprehension). The procedures for conducting IRIs are quite traditional as they are described in the material that follows. Teachers should, however, consider the content and the students they teach, not necessarily insisting on the exact steps. Rather, adjustments should be made in according to the material that is being taught and the age of the student.

### *Levels of Performance*

Four different levels of performance may be considered, corresponding to the ease with which the student reads, reflecting performance on both reading and reacting.

**Independent reading level.** This level is one in which "children can function on their own and do a virtually perfect job of responding to the printed material" (Johnson, et al., p. 13). Quantitative guidelines for this level include reading with 99% accuracy of word recognition and 90% accurate performance on the comprehension tests. Both criteria for word recognition and comprehension must be met for independent levels to be determined.

**Instructional reading level.** This level is one in which children can be meaningfully taught. Quantitative guidelines include oral reading with 95% accuracy, and performing with 75% accuracy on the comprehension component. A range of different instructional levels is more likely to be found with most children. A child may be performing at a different level in natural science than she is in a social science or in the basal reader.

**Frustration reading level.** This level is defined as one in which "the child becomes completely unable to handle reading materials..." and is assumed to be frustrated (Johnson, et al., 1987, p.19). The suggested quantitative criteria for this level include oral reading with 90% or less accuracy and performing at 50% or less accuracy on the comprehension component.

**Listening comprehension level.** This level is defined as the highest level at which children can satisfactorily understand materials. The quantitative summary at this level includes performing at 75% accuracy on comprehension questions asked about the materials (Johnson, et al., p. 20).

### *Administration Procedures*

Procedures for administering, recording, and scoring individual informal reading inventories are relatively straightforward. Appropriate passages are selected, with the guiding question being whether the selection seems similar to most other selections included in this book. Typically, two or three selections are sampled at each level: one for oral reading, one for silent reading, and the third for evaluating listening comprehension. Generally, the passages are of increasing length from pre-primer through the highest level: Pre-primer to primer has 50-75 words, first reader to second reader has 100 words, third reader to fourth reader has 150 words, fifth reader to sixth reader has 200 words, seventh reader and above would have 250 words. Following selection of materials, the remaining steps for administering a complete IRI have been delineated by Johnson, Kress, and Pikulski (1987) and described by Tindal and Marston (1990).

**Step 1. Preparation for testing.** Prior to testing, materials should be collected and testing procedures identified.

**Step 2. Establishment of rapport.** Students should be told why they are being tested and what they will have to do.

**Step 3. Determination of the level at which to begin.** Students should be given materials that are appropriate based on reviews of school records and folders and previous teachers' judgments.

**Step 4. Establishment of readiness and purpose for reading.** Student s should be given a reason for reading (e.g., "I would like you to read this selection on mountain climbing to find out what you can about this topic").

**Step 5. Oral reading.** While students are reading, all errors and other features of oral reading should be tracked. In the oral reading component of IRIs four types of errors typically are counted: mispronunciations, insertions, omissions, and requests for examiner aid (or hesitations).

**Step 6. Assessment of comprehension with orally read materials.** Immediately after the oral reading of a passage, students should be asked a series of questions that have been prepared ahead of time. In this step, a student is not typically given access to the reading materials, but must answer the questions or describe the content as best he/she can based on his/her oral reading. At least 10 questions should be prepared and should be context-dependent so that they cannot be answered without reference to information in the passage. The types of questions described in the section on

content knowledge could be considered here (facts, concepts, principles along with reiteration, summarization, illustration, prediction, evaluation, and application).

**Step 7. Silent reading.** Students should be observed silently reading a second selection that has been sampled from approximately the same level.

**Step 8. Assessment of comprehension of silently-read materials.** The same strategy used for comprehension with oral reading should be employed: Ten questions are asked, dealing with factual, inferential, vocabulary, and evaluation issues.

**Step 9. Oral rereading.** The examiner should establish some purpose for orally rereading a portion of the selection that previously had been read in silence. The main purpose here is to assess the student's skill at skimming, measure ability to read for a specific purpose, and determine the difference in fluency having read the material previously .

**Step 10. Test in different materials.** Students' performances on both oral reading and comprehension should be assessed on either more or less difficult material.

Of course, these 10 steps represent a very thorough administration of an IRI, and frequently a more abbreviated format is actually employed. Great attention often is devoted to steps 5 through 8, in which students orally and silently read and are asked comprehension questions.

## Active Academic Engagement

Another strategy for determining whether students are placed in an appropriate instructional level is to observe their engagement and analyze the products they create. The idea of using engaged time as a measure for placement actually has its roots in many different educational research efforts.

Back in 1963, John Carroll wrote one of the most influential articles in education, entitled, *A model of school learning*. He described a model that focuses on time as the prime measure for determining an individual's success in school. Students need time to learn. The amount of time needed is influenced by such things as learner aptitude, ability to understand instruction, and quality of instruction. Students also spend time learning. The amount of time spent is influenced by such things as opportunity to learn and willingness to spend time learning. Thus, the complete model postulates that the degree of learning is a function of the *time actually spent learning* relative to the *time needed to learn.* Instructional level becomes important because of the influence from any of the three "within-learner" variables (aptitude, ability to understand instruction, and perseverance). When students are placed inappropriately, we will see a mismatch with any or or all of these three variables.

Although this conception of time as a major influential variable in learning was identified in the early 1960s, it was 15 years later before any actual research was begun. In the late 1970s, Far West Laboratories embarked on a series of studies to ascertain the beginning skills of teachers. The project came to be known as the *Beginning Teacher Evaluation Study*, which is really a misnomer. They quickly moved the focus of research toward teacher behaviors that appeared highly correlated with increased achievement. Adopting Carroll's' model of school learning, they found that a very high correlate of achievement was active learning time, which they defined as the amount of time a student spends actively and successfully (above 90% accuracy) interacting with material. From this initial research came a slew of studies on time engaged in learning, its correlates, and outcomes (Denham & Lieberman, 1980; Fisher & Berliner, 1985; Graden, Thurlow, & Ysseldyke, 1982). Quite consistently, regardless of how time is defined, the amount of time that students spend actively and academically interacting (with materials or individuals) is highly related to achievement.

So, how does engaged time relate to appropriate instructional level? For instruction to address assessed levels, students must be appropriately engaged and interacting with material or individuals. This engagement may be a function of any of the variables listed above, whether they are internal to the learner (special talents or aptitudes, general ability to understand instructions, or perseverance) or arise from external conditions, including their opportunity to learn or the quality of instruction. Unlike the previous system, in which placement was static and occurred prior to instruction, this system is dynamic and occurs throughout instruction.

Measurement quite simply focuses on the critical effect, which is engagement in learning and is expressed as time. When no time is spent in learning, we can assume the student has not been placed in an appropriate level. When the student is engaged in learning a great proportion of the time, we can assume that the s/he has been placed in an appropriate level. Measurement of engagement, however, must be clearly operationalized. It is not the amount of time that the teacher planned to teach. Nor is engaged time the amount of time allocated for instruction. Finally, it is not the amount of time the student spent in instruction. All three of these definitions miss an important element of the definition: active interaction that is successful.

## Case Example

A case example may help illustrate the use of these two measures of assessment for instructional level. In a program for a group of 2 fifth-grade students, the teacher has developed daily enrichment activities centered around science. Each student is proceeding in a different area. One student is studying the human body (its skeletal and organ systems), and the other student is studying outer space. After presenting organized materials and activities, the teacher starts them

on a reading and reacting course of study; later the students will work on special projects and engage in some field trips.

The first task is to find some appropriate material for them to read. After a visit to the library (in school, downtown, or at a local university), the teacher brings in several books that appear to be appropriate. However, at this point, we may not be sure. Therefore, the assessment process should at least begin with determining the students' reading skills. Some sample passages are selected and each student is asked to read aloud. The quality of this reading is assessed formally (using an error count) and informally (listening to the prosodic features of the student's reading). After a certain amount of material or time, the teacher asks the student to continue independently. After the reading, the student is asked a number of questions, some of which involve factual reiterations of the material and others which ask about the meaning of various concepts. The results of the first (oral reading) component can reveal either accuracy or rate as well as a judgment of quality. The results of the second component, answering questions, can reveal student reactions to the material. Both pieces of information can be used to determine if the reading material is appropriate. Other passages can be sampled if the teacher wants to confirm this assessment or obtain a more representative sample. Throughout the delivery of instruction, this procedure may be reinstated to check on new materials.

After the students have been placed in their respective units of study (the human body and outer space), the teacher observes them carefully, noting the amount of time they are actively and successfully engaged. This procedure may be accomplished more or less formally. On the informal end of the continuum, the teacher simply makes a subjective judgment about the percentage of time they spent actively interacting with the material or the teacher during the lesson. On the more formal end, the teacher can take "engagement checks" periodically throughout the period, marking down on a card whether they are engaged or not engaged. At the end of the period, they can total the number of times they saw the student engaged, divide it by the number of times they observed, and obtain a percentage. For example if the teacher observed the student 7 times during the hour (on the average, every 8 minutes an observation was made) and found that they were engaged 5 of the 7 times, we could state the student was engaged about 70% of the time.

The purpose of these measures is simply to ascertain whether a student is working with material and or content successfully and interactively. If many mistakes are made and the accuracy of a student's work is low, the student has not been placed appropriately. If the student is not interactively engaged with the material, appropriate placement is equally threatened. In both cases, we know less about why the accuracy or engagement is low, only that is is low. It could be a function of student motivation (perseverance), interests (aptitude), quality of instruction, etc. The important decision to make, if either accuracy or engagement is low, is whether to change the program somehow: Introduce different materials; implement different instructional procedures; modify the emphasis from book learning to experiential learning; interact differently with students; and, most importantly, continue to assess their success after the introduction of any instructional changes.

## Assessing Learning Rates

In Caroll's conception, time is the critical dimension. By tracking engaged time, we have an idea of the amount of time spent learning. However, we also need to know if the material has been learned. The formula states that learning is a function of *time spent actually learning relative to the time needed to learn.* The numerator (time actually spent learning) will be the smallest amount that arises from any of the following three factors: (a) opportunity, (b) perseverance, or (c) aptitude. The denominator (time needed to learn) is determined by (a) aptitude, (b) the quality of instruction, and (c) ability to understand instruction. The ideal outcome would be a ratio of 1/1, that is, the time spent in learning is equal to the time needed to learn.

Yet, assessment of learning rates also must address the content of what is learned. Such content can potentially influence all of the factors noted above: opportunity, perseverance, aptitude, quality of instruction, and ability to understand instruction. Which specific outcomes should be included? Most of the material in the first 50 pages (dealing with basic skills, content information, and procedural knowledge) focused on performance outcomes.

Furthermore, a conception of learning must include judgments for determining whether learning indeed has taken place. How should performance be judged? When can we be certain that the skills, content, or routines are really learned? Are there any standards we can use to help us make this discrimination?

When considered in conjunction with time spent learning (or time needed to learn), both of the last two issues, performance outcomes and standards for making judgments, provide the necessary components for ascertaining learning rates. However, whether we are using interactive observations, permanent products, or tests/measures, we can only sample a finite amount of learning. We simply cannot assess students on everything we taught them or everything they know. Hopefully we anticipate that the performance outcomes on our assessments accurately and adequately reflect all the outcomes that are possible in the domain. In moving Carroll's conceptual model to actual practice in the classroom, therefore, we must develop a system for sampling specific instructional content (basic skills, content information, or procedural routines) and inter-

preting performance according to some standards.

One final issue also must be resolved: the degree to which instruction and assessment are related. In teaching a lesson, many cues exist that students "understand" the content. These cues can be used to pace the lesson and make decisions about reviewing the material. However, our record keeping is often quite inconsistent and possibly inaccurate. For example, within an instructional episode, teachers may ask questions closely related to the content of instruction. If they ask a student a question, and the answer is correct, can we say that learning has occurred? To answer this question we may need to consider the relationship between the instructional and assessment content. When an extremely close relationship exits, we are generally less certain that learning has occurred. In contrast, when a distant but related relationship occurs, we are generally more certain.

I will use two evaluative systems that help delimit specific performance outcomes and allow us to devise standards for interpreting them. The first is a criterion-referenced approach that is based on mastery of separate units, and the second is an individual-referenced approach that is based on improvement over time. These two evaluative systems represent different approaches to sampling behavior and interpreting whether learning occurred. When we add in the element of time, we also can ascertain the pace at which learning has occurred.

In a *criterion-referenced evaluation*, the focus of assessment is clearly on *what* the student actually can or cannot do on specific skills and knowledge tasks; frequently performance is interpreted in terms of *mastery*, which connotes a judgment that learning has occurred. Synonyms may be *proficiency, fluency, facility*, etc. The term can be used in many different arenas of learning and is not limited to academic tasks. Importantly, the term implies an absolute standard of acceptance. In this system, we focus on the rate at which different materials (different units, concepts, chapters, etc.) are being learned. Absolute criteria are used to make judgments: Interpretation is made in reference to a specific level of performance on a scale that is noncontinuous (all levels of performance above this cut-off are considered mastered and all levels below it are considered non-mastered. The differences between scores within either side of this cut-off are less critical than those across the cut-off.

Rather than using mastery, which is based on an absolute cut-off, an *individual-referenced* evaluation reflects change in material that is comparable over time (material is alternately equal). No specific performance level is identified (above which there is mastery and below which there is non-mastery); rather, change is noted on a continuous scale, from less to more, noting the direction of change over time and the rate at which it is changing. The materials are not qualitatively different and alternate forms are used to generate comparability across assessment tasks.

Both systems should be considered as guides for assessing learning rates. They allow interpretations to be made for judging performance outcomes. For example, if you were told that a student in your class had received 35 points from an assessment (as the result of an interactive observation, a permanent product, or a test/measure), and were told nothing else, how would you interpret this score? Does the number 35 mean anything to you? Probably not. Even if you were told that this score was attained on a math story-problem test with 50 possible points on it, you wouldn't be in any position to interpret the student's performance. The following section focuses on interpretation of student performance.

### *Establishing Evaluative Standards*

Basically we have three strategies available for interpreting of student performance: norm-referenced, criterion-referenced, and individual-referenced. Each strategy provides an interpretive guide that helps you understand and appreciate a student's performance. For the moment, you should consider these three strategies as quite distinct. However, they can be intermixed somewhat, providing us a blend of interpretive guidelines.

In a *norm-referenced* approach, the student's performance is compared to other students (who are comparable in age, cultural background, race, sex, etc.). The important index is how the student compares to them—their relative standing. In the example above, the score of 35 is very interpretable if we had also been told that the average for the group was 30 and the average amount of variation (standard deviation) was 5. With this information we can say that the student is above average; in fact, she or he is at the 84th percentile. Not bad, maybe. In a norm-referenced approach, a number of different metrics are available. Nonetheless, they all interpret performance as relative group standing.

In a *criterion-referenced* approach, we are not concerned with the student's standing in a group, but with performance on well-defined tasks. In the example above, we are more interested that the items used to develop the test represented a random sample of single-operation (using addition, subtraction, and multiplication) math story problems from the first half of the math book. The score of 35 out of 50 is somewhat more interpretable: On these kinds of problems, our student answered 75 percent of them correct. In a criterion-referenced approach, we must always define the domain, including how we sampled the items. Notice that the problems had been randomly sampled (in fact, we could calculate the percentage of problems this sample covers). Generally, we also use some guidelines for determining success, often referred to as mastery or proficiency. In most curricula tests/measures, the cut-off score for defining mastery from non-mastery is

somewhere between 75 to 90 percent correct; however, this cutoff can be established anywhere.

In an *individual-referenced* approach, we interpret performance by comparing the student's score to previous performance levels attained by that student. Rather than compare performance to other students or to some absolute standards of mastery, the important dimension is whether an individual's performance has improved. This strategy is very much like the stock market's Dow Jones Average, which increases or decreases relative to the previous day's performance. In our earlier example, the score of 35 would mean something if we had been told that on previous weekly measures that sampled similar (not the exact same problems), the student had scored 20, 25, 21, and 30. With this information, we can interpret the score of 35 as definite improvement. Furthermore, we can see improvement has been occurring quite consistently for the past month.

All student performance can be interpreted according to these three guides. However, they do carry with them different assumptions, and each has its own advantages and disadvantages. I will focus on only the second two guides, since a norm-referenced approach is generally inappropriate for evaluating change in performance. For other decisions, like screening and placing students in specialized programs, a norm-referenced approach is about the only possible strategy. However, for our purposes, they have the following major problems:

1. The items represented on the measure are broadly sampled because they have to accommodate students from a wide range of skill levels; very few items are instructionally useful.

2. Growth is difficult to ascertain, particularly at the extremes (very low and very high scores). Of course, improvement on individual items can occur (i.e. the student can get a raw score gain from the first to the second testing). However, since we use relative scores rather than raw scores to interpret change, it is likely that this raw score gain disappears when converted. That is, the student may perform at the 98th percentile at time 1 and at the 98th percentile again at time 2. This outcome means that, although the student probably answered more items correctly, his/her relative standing in the group did not change from time 1 to time 2.

3. Since a major purpose of assessment in this monograph is to ascertain rate of progress, the assessment system needs to have many alternate forms that can be implemented frequently in the classroom. Norm-referenced measures, at best, have only two alternate forms, and therefore can be used only for pre- and post-testing. Although they may appear to be useful for placing students into instructional groups and/or materials, you should be forewarned: Norm-referenced measures are as likely to over- as under-place students into curricula; they are marginally useful for instructional (ability) grouping students.

In summary, only two options are available for ascertaining a student's learning rate: criterion-referenced and individual-referenced evaluations. No more information will be presented about norm-referenced measures. The three assessment methodologies are more-or-less applicable for both criterion-referenced and individual approaches. As mentioned earlier, however, they each have some assumptions embedded in them, and they have advantages and disadvantages. In the material that follows, I will examine each approach, consider implications from its use, and offer strategies for displaying results.

## Criterion-Referenced Evaluation

Criterion-referenced evaluation closely matches assessment with instruction; it is predominantly used in the classroom and most curricula (i.e. model-lead-test). I look at it as the "near-sighted" approach: Focus is directed upon objects within a short distance, while objects that are far away are unfocused.

### *Procedures*

To implement this evaluation strategy, three steps have to be completed:

1. A domain of items must be defined. This domain can be included within interactive observations, permanent products, or tests/measures. It must, however, be clear in specifying the boundaries of the domain. Which item types are in and which item types are out? For example, the following domains are all clearly specified: addition math facts 1-9; spelling words with a consonant-vowel-consonant (CVC) construction, concepts from the instructional unit on the Civil War (in U.S. History, chapter 9), principles relating velocity and force. In all of these examples, the skill or knowledge content is quite clearly defined. Not all assessment domains are as clear, as depicted in the following examples: fourth-grade math problems, reading vocabulary words, geography of the Far East, the biomechanics of movement. These domains only provide a general picture of the instructional (and assessment) content. Interpretation of performance in a criterion-referenced approach should be as much a part of the domain definition as it is with the score or outcome performance level. When you see the outcome, you should be able to understand exactly what was taught and learned.

2. Once a domain is defined, some strategy needs to be developed for sampling items from that domain. Generally not all items within the domain can be sampled; therefore, we have to come up with a system for selecting only some of them. This system should be sufficient to give us confidence that, although the student did not answer all items, he probably would have answered them correctly.

An obvious example could be constructed from our CVC domain above. Do we need to present all 100

words to determine whether the student can sound out or spell the different consonants and vowels? Probably not. Instead, we simply could take a random sample, present it to the student, and if performance is adequate, assume that performance on items not presented would have been the same. While several different approaches (too technical to cover here) are possible, you should note that the item selection process can influence the quality of the assessment, particularly with interactive observations and tests/measures; it is uncertain how the sampling system influences permanent product analysis.

3. This last step is optional, though typically included in most criterion-referenced assessments: determination of mastery or proficiency. Some level of performance is set, above which student performance is deemed adequate, acceptable, proficient, etc. and below which it is deemed inadequate, unacceptable, non-proficient, etc. This step, if employed, is fraught with controversy. Many journal articles and entire books have been written on this subject, and given the complexity of this issue, you may want to look at a measurement book to better understand what it means to use a mastery cut-off.

The only two issues that we will consider here involve the number of items and the certainty of the decision. Generally, more items are better than fewer items, particularly with reference to a mastery measure. Although many tests/measures have as few as one or two items for making mastery decisions, this number is probably too few. Generally, a minimum of 10 items are needed to be certain of mastery. Of course, the number of items that are needed is quite dependent upon the definition of the domain. With very highly constricted domains, fewer items are needed; with broad domains, many more items are needed to be certain that all those not included on the assessment are mastered. The last issue, certainty of the mastery decision, can best be described as follows. Each mastery measure has three zones: clear mastery, uncertain mastery/non-mastery, and clear non-mastery. The outside two zones (clear mastery and non-mastery) generally present no problem. The dilemma appears with that indifference zone (Shepard, 1984), where there is uncertainty about proficiency. I recommend that you include all three outcomes in summarizing performance.

### *Implications*

The advantage to a criterion-referenced evaluation is the proximity between instruction and assessment: They are well aligned. To borrow a phrase from personal computer jargon, "What You See Is What You Get" (WYSIWYG). Furthermore, it is relatively easy to implement, and it follows an orderly four–step process:

1. Instructional goals are defined, in which materials and activities are specified.
2. Instruction is implemented
3. An assessment is completed. The materials and activities specified in step 2 provide the domain for sampling items.
4. A decision is made: Is performance sufficient to call it mastered or acceptable? If so, new materials and activities are organized for another set of instructional episodes.

And on it goes. With each new set of materials and activities, new assessments are constructed.

You should be aware of two serious drawbacks of this system. The most obvious problem, as mentioned earlier, is the difficulty in establishing mastery or proficiency. No clean and proven technology exists yet to eliminate the need for caution. Be sure to include enough items or samples of behavior, and consider all three zones when making decisions. The other problem, though less noticeable, relates to the "near-sightedness" of criterion-referenced evaluations. Since instruction and assessment are closely matched, and assessment does not occur until instruction has been delivered, two types of error can be made. First, the student is paced no faster than instruction and assessment (no preview performance levels are ascertained). Second, retention is assumed, since each assessment closely focuses only on the material that was taught (no review of performance is ascertained). Both problems can be overcome only by systematically building in preview and review assessments, which are really incorporated into the long range sampling strategy of the individual-referenced evaluations.

### *Displaying Performance Outcomes*

In tracking performance with a criterion-referenced evaluation, you should consider the non-comparability of the assessments. Because each assessment is on qualitatively different material, comparison of performance across outcomes is confounded. For example, in an instructional unit on "peoples of the world," a student could master the information presented on subcultures in the U.S. but be non-proficient with the Kurdish peoples of Turkey. However, since these two units are so different in materials, any performance scores will be unrelated. It is also quite likely that the kinds of items (facts, concepts, and principles), as well as the sampling strategies, are also different. Therefore, the only summary of performance that is possible in a criterion-referenced evaluation is the mastery status itself.

Two record-keeping systems can be developed to record mastery: the typical classroom gradebook and the graph. The gradebook should note the following information, the first two of which are often missing:

1. A phrase description of the instructional and assessment content and the dates inclusive of its coverage.
2. The type of assessment that was conducted (i.e. an interactive observation, analysis of a permanent product, or test/measure).

3. The number of opportunities presented or total possible score attainable.

4. The actual score of the student, with some system for noting whether this score is above the mastery or proficiency level (e.g., by circling the number attained).

All students are listed in successive columns. By listing students in columns (and content in rows), more descriptive information can be included on the assessment process. This system should accommodate at least 3-5 students on a standard sheet of paper.

The second strategy for depicting performance is to graph the mastery progress of the student. The graph should employ the following conventions:

1. Draw a vertical and horizontal line so they intersect at 0. Label the vertical axis with the title "Point Totals and Mastery Status." The horizontal axis can use either successive numbers (1 to N) to represent chapters or informational units or the phrase descriptor of the content (and type of assessment methodology), all of which can be written below the axis. In labeling the successive units, begin at the left and count to the right, since that is the order of presentation. Finally, label the horizontal axis with dates that mastery was assessed (use real time and either school- or week-days).

2. Student data points (representing mastery or proficiency) are recorded as different symbols and a line is used to connect them over time. The only rule to follow is to move over and up when there is mastery and over (but not up) when there is non-mastery. Since this graph communicates all of the same information as the gradebook, except the number of items or total score possible, this number may be placed in parentheses below the data symbol.

3. Successive values are recorded on the graph whenever an assessment is conducted.

An example of each type of record is displayed in the two figures below. Table 1 employs a gradebook and Figure 5 employs a graph. Later, after presenting the individual-referenced evaluation system, the two decisions of placement and ascertaining learning rates are reviewed using these two systems.

Table 1. Gradebook Illustrating Criterion-Referenced Evaluation

| Instructional Content-Dates | Type of Assessment/ Number of Opportunities | John's Performance | Sarah's Performance |
|---|---|---|---|
| Geology Concepts-12/5 | Interactive Observations | | |
| Bedrock | 5 | 3 | 5 |
| Continental Drift | 5 | 4 | 3 |
| Continental Shelf | 5 | 5 | 4 |
| Epicenter | 5 | 5 | 5 |
| Earthquakes | 5 | 5 | 4 |
| Richter Scale | 5 | 5 | 5 |
| Chemistry Concepts-1/4 | Test/ Measure | | |
| Atom | 3 | 3 | 3 |
| Atomic Number | 5 | 4 | 5 |
| Atomic Weight | 5 | 5 | 5 |
| Half Life | 3 | 2 | 3 |
| Chemistry Principles-1/4 | | | |
| Fission | 3 | 2 | 3 |
| Fusion | 3 | 2 | 3 |

## Individual-Referenced Evaluations

This evaluation system focuses on change over time, with rate of improvement as the primary datum for determining if programs are working. As mentioned earlier, this system is the "Dow Jones Average" of education. As in the stock market, the two important characteristics involve a historical look at how much improvement has occurred in a certain time period and a prediction about future rates. If you find that stock in which you just invested is beginning to rise, you are likely to hold on to it, at least in the short run. In contrast, if it begins to drop, you may bail out while you can. Although this system is conceptually easy to understand, it has certain technical characteristics that are quite subtle, but very important. In the following material I will first describe some critical features, then outline the implementation procedures. Next, I will present an analysis that includes both advantages and disadvantages, followed by a system for recording and communicating assessment results.

To compare a student's performance to previous levels and eventually calculate the rate at which learning is occurring requires comparability between all the data values. A criterion-referenced test or measure could not be used in this manner since each measure is a unique opportunity and is not directly comparable to other raw score values. Therefore, the sampling plan needs to be broader than that used for a criterion-referenced evaluation; it has typically been described as long range goal assessment. Rather than matching the content of assessment with that of *an* instructional episode, the domain for assessment spans *several* instructional episodes.

As I said earlier, the major limitation of a criterion-referenced evaluation is "nearsightedness," and the results from any one assessment have little generalizability over time or across items. In contrast, the sampling plan of an individually-referenced evaluation has built into it both a preview and review component. This feature, however, makes it unlikely for content information assessment. It is best employed with either basic skills or procedural knowledge. Individual-referenced evaluations imply a certain kind of uniformity in the knowledge or skill being assessed, so that both near examples (closely related to instruction) and far examples (appropriate generalizations not directly taught) can be included within any given assess-

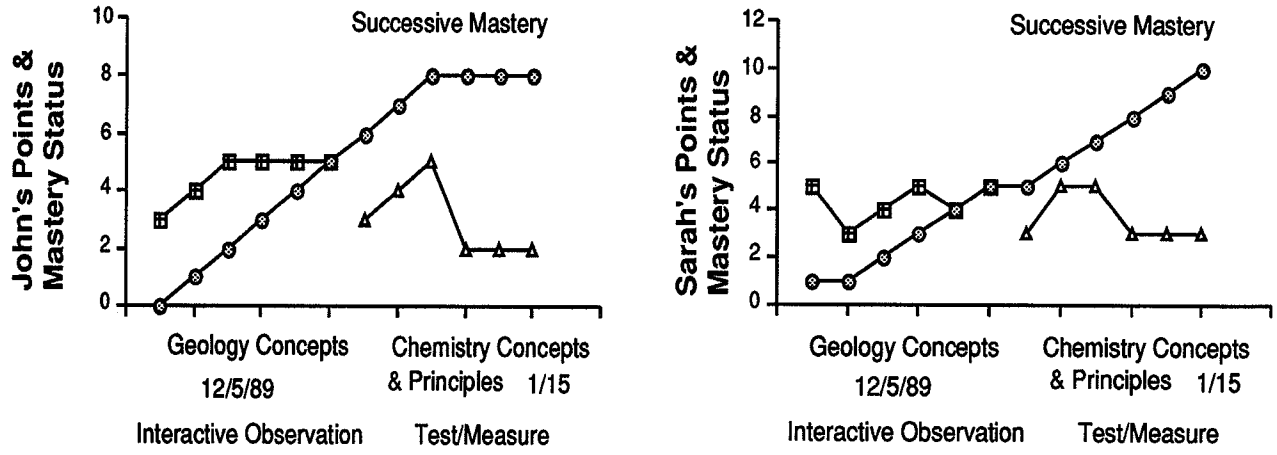


Figure 5. Graphs Illustrating Criterion-Referenced Evaluation

ment task. Since content information is so specific, and probably has few far examples that represent generalizations not directly taught, this form of evaluation consequently is restricted.

The following examples all represent an appropriate focus on generalized learning of basic skills and procedural knowledge. In spelling, many words follow phonetic rules quite consistently (not perfectly). We can get a picture of the student's proficiency by including representative items from these various word families; we certainly don't have to include all possible items. Reading words (decoding them) also provides many examples in which various structural characteristics can be considered, with a measure including a representative sample of them. Math computation problems are probably the most lawful in terms of generalizations. In fact, they can be considered as lawful with no exceptions to the rules. Most basic skills, as we have defined them, lend themselves well to this form of assessment. Procedural knowledge also can utilize an individual-referenced evaluation. Here, the generalizations involve rules across different problem types. Most math story problems and science problems include routines that generalize past the immediate items included on an assessment. Although the wording may change, and the contexts may vary, the routines, nevertheless, remain constant. Continuous measurement is possible, since comparability is ensured in the items.

### *Procedures*

The following steps can be followed in conducting an individual-referenced evaluation:

1. A domain needs to be established that includes all material and activities from the entire instructional series, from now until the end of instruction (i.e. a long range goal for the end of the year). For example, a teacher may indicate that the students, who are now reading at the beginning of a certain book, should have finished with it by the end of the school year. This material represents the long range goal. Likewise, a math teacher may decide that one-step story problems with addition and subtraction represent the long range goal for a student.

2. Alternate assessment samples are then devised by selecting representative items. For example, a teacher may randomly select passages that are typical of most others in the book and have students read from them, asking them to retell the story after each reading. Or a series of math problems may be collected that typify the type of problems in which the student will come in contact during instruction. It is this step that precludes the use of content information for individual-referenced evaluations.

3. The assessments then are collected on a schedule—every week, biweekly, or monthly. In order to develop an adequate database for viewing change in performance, a certain number of data values (7 to 12) should be generated.

In summary, an individual-referenced evaluation uses time-series information to ascertain not only how much improvement is occurring, but also the rate of this improvement. The student's performance is interpreted by looking at previous levels (over time). If a program is working, you should see a general trend of increased performance; if the program is not working, such an increase is not discernible. To evaluate program effects, the general trend of improvement (referred to as slope) and the amount of variation or fluctuation (referred to as variability) should be used.

### *Implications*

An individual-referenced evaluation employs the ultimate criterion in education: Are students retaining the necessary information to solve a wider range of problems, some of which may be generalizations of strategies, and with greater automaticity or fluency? Indeed, the programs we devise for students are not really meant to increase a student's relative standing, which is the outcome of a norm-referenced evaluation. Nor can we be certain that an ill-defined state referred to as mastery, reflected on a list of specific skills, devel-

oped from a criterion-referenced evaluation, is the end goal of education. However, we can be certain that, with an individual-referenced evaluation having a positive trend, improvement has occurred, the program is having an effect, and that such improvement is meaningful. The student's performance is not being compared to that which has been attained by others or interpreted using a list of skills having an absolute level of performance required; rather, the student is being compared to himself or herself.

Because assessment proceeds from a goal-oriented view, which is far-sighted, the items represent both preview and review. The content of instruction and the content of assessment are not tightly linked, as in a criterion-referenced evaluation. This strategy, therefore, provides a valuable adjunct to the immediate outcomes from instruction, reflecting appropriate generalizations that have been acquired by the student. For example, if a student was just taught the rule for doubling a consonant when adding a suffix, and then is immediately tested, we might not be very confident that even 100 percent accuracy reflects substantive learning. However, if we have taught the rule, along with other spelling rules, and include items that incorporate many different rules intermittently over many different occasions (alternate forms), we can be more certain that the rule has been learned, if it is correct on the assessment.

An important part of this far-sighted approach is that any one data value is quite limited in making interpretations. Since the domain for item sampling is broad and the strategy for actually sampling items is often random (within some strata), any one assessment may have a disproportionate number of items of a certain type. For example, in the reading example above, an easy story may have been selected; or the story may have just been covered in class. In either case, the data for that day may be very high. The next measure may sample a more difficult passage or one that has not been taught. Consequently, performance levels are much lower. Therefore, to appropriately utilize this approach, the general trend of data must be considered, not any specific data value in isolation. If the student is becoming a better reader, the number of words read correctly will generally improve over time.

A word of caution is in order, as this approach is not without its drawbacks. Such individual-referenced evaluations must be interpreted in relation to the material used for assessment. A very steep slope may indicate that learning is rapid and there is a ceiling effect, in which performance improvement will no longer be visible (the maximum level has been attained). In contrast, a very low slope, or no slope, may actually reflect assessment materials that are too difficult and are insensitive to change. A floor effect has been attained and a considerable amount of time is needed before any discernible changes appear.

Several other limitations also should be noted. The definition of the domain from which items are sampled can be a major potential problem. Often it is difficult to project into the future. Few standards exist for determining an appropriate amount of material to sample for assessment. Short of applying this strategy and becoming familiar with it, my only sage advice is to pay attention to realistic long-range goals. As discussed below, guidelines for evaluating instruction in this system are quite flexible. In fact, few firm procedures have been identified, except to increase learning rates as much as possible. Finally, domains from content information are probably not appropriate; the system is appropriate for either basic skills or procedural knowledge, where alternate forms reflecting generalizations across item types are embedded within the assessments.

### *Displaying Performance Outcomes*

As in the criterion-referenced evaluation, two different record-keeping and communication systems may be employed: a gradebook and a graph. Clearly, a graph is preferred, although both are presented below. I will first delineate the steps in developing them and then present completed examples.

In developing the gradebook system, three components should be included: (a) a description of the assessment materials (where they were sampled from and how they were sampled), (b) columns that list the date and the score, and (c) the units for scoring performance.

This information can easily be transferred to a graph to more quickly reflect the results. The graph is constructed in the same manner used with a criterion-referenced strategy: Two axes are plotted, with the vertical one depicting the performance scores and the horizontal one depicting time (successive days or weeks). Both axes should be clearly labeled and include appropriate values or dates. At the top of the graph is a description of the student performance outcomes and in the graph itself are data values plotting the actual scores as correct and/or incorrect. The only remaining rule to remember is to connect successive values together with a line, with only two exceptions: If a large break occurs in time and no data have been collected or if an intervention change has been implemented and is reflected in the graph by a vertical line. In both cases, the line connecting successive data values is broken and re-established after the data collection is continued or after the vertical line.

Since I have implied that a picture is worth a thousand words, a couple more figures are in order. In Table 2, a gradebook system is depicted; in Figure 6, a graph is drawn for each student. The biggest advantage to the graphs in Figure 6 is that the two major indices of time series data—slope and variability—can be physically drawn into the picture. In this figure, a line of best fit can be superimposed across the successive data values. This line should reflect the general trend in performance; if we want to project perfor-

Table 2. Gradebook Illustrating Individual-Referenced Evaluation

| Random Sample of Passages from Basal Reader | Dates Assessed Using Tests/Measures | Jane's Oral Reading: Correct & Incorrect | Susan's Oral Reading: Correct & Incorrect |
|---|---|---|---|
| Page 105 | Dec. 1, 1989 | 123 Cor/1 Inc. | 103 Cor/ 1 Inc. |
| Page 210 | Dec. 5, 1989 | 113 Cor/ 2 Inc. | 93 Cor/ 3 Inc. |
| Page 190 | Dec. 8, 1989 | 139 Cor/ 1 Inc. | 113 Cor/ 5 Inc. |
| Page 35 | Dec. 11, 1989 | 102 Cor/ 3 Inc. | 123 Cor/ 3 Inc. |
| Page 240 | Dec. 13, 1989 | 134 Cor/ 5 Inc. | 128 Cor/ 5 Inc. |
| Page 113 | Jan. 5, 1990 | 149 Cor/ 8 Inc. | 119 Cor/ 1 Inc. |
| Page 188 | Jan. 9, 1990 | 140 Cor/ 2 Inc. | 130 Cor/ 2 Inc. |
| Page 56 | Jan. 12, 1990 | 147 Cor/ 1 Inc. | 125 Cor/ 1 Inc. |

mance levels in the future, we can simply extend it outward and extrapolate a value for a certain date. For example, in the first graph, it appears likely that in two months the student will be performing at a 175 word correct on the assessment. For practical purposes, you should try to draw a line that seems to reflect the general trend. It should pass through the middle of the data array so that most values are equidistant above and below it and so that it follows the general contour of their change over time. More accurate and sophisticated systems can be learned, using either hand-drawn or computer-generated slopes (see Tindal and Marston, 1990, for a full description of the options). In like manner, variability of performance can also be drawn over the data array. To do this, simply draw a line through the value furthest above the slope and parallel to it. Draw a similar line through the value furthest below the slope and parallel to it. The band of values included within this envelope reflect the amount of variation present in the time series.

In both graphs, a slight improvement is evident, with the general trend moving toward more fluent reading. The improvement is even more impressive, given the time of the year (Christmas break). And, for both students, there is only slight variability in their day-to-day performance.

## Summary

Great flexibility exists for assessing students. Rather than viewing this process as an extra burden, you can think of it as an opportunity to develop effective instructional routines and a database for confirming them. Teachers need to be actively involved in structuring appropriate assessment tasks for placing students into instructional levels and ascertaining their learning rates. Given the differences among teachers, assessment practices are likely to follow in kind. Therefore, in developing a database, teachers should first define the purpose of instruction: What are the performance outcomes that you would be satisfied with as representative of the material taught. Then, develop some system for actually collecting this information. Finally, the decision itself needs to be made, using either of two evaluation standards (criterion- and individual-referenced). Implicit in this process is the use of data to inform decisions. By using the graphic summaries of student performance, that decision-making process is likely to be both more expedient and more effective.

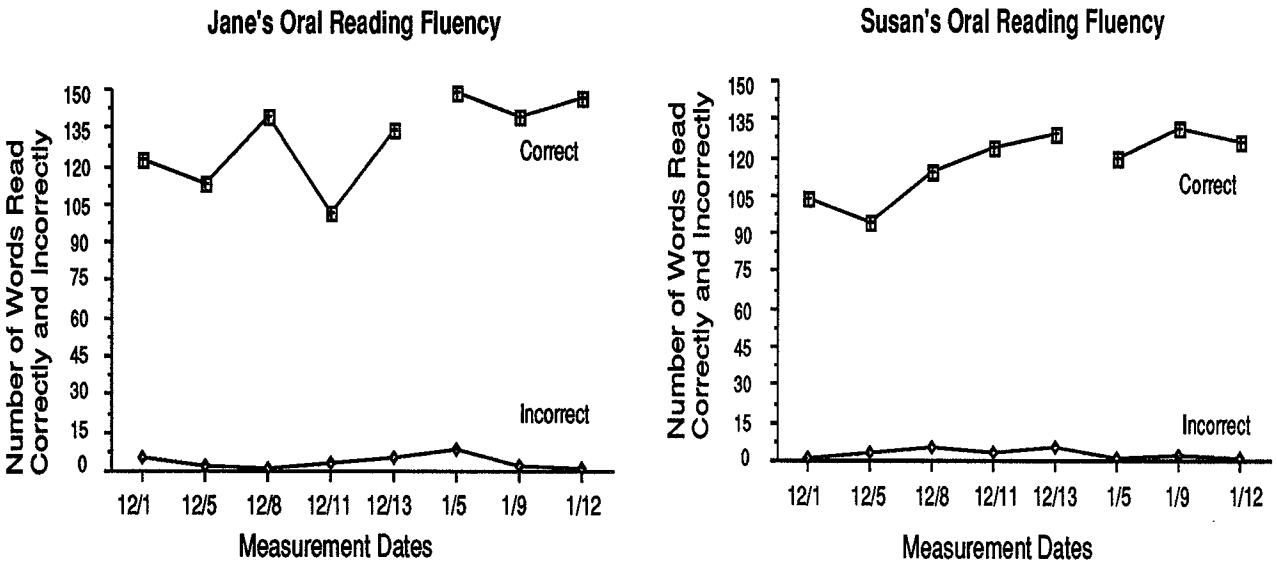


Figure 6. Graphs Illustrating Individual-Referenced Evaluation

## References


Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to improve learning.* New York: Hill Book Company.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. The classification of educational goals - Handbook: The cognitive domain.* New York: David McKay Co.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723-733.

Denham, C., & Lieberman, A. (1980). *Time to learn.* Washington, D.C.: National Institute of Education.

Fisher, C. W., & Berliner, D. C. (1985). *Perspectives on instructional time.* New York: Longman.

Gagne, R. (1985). *The conditions of learning and theory of instruction.* New York: Holt, Rinehart, and Winston.

Graden, J., Thurlow, M., & Ysseldyke, J. (1982). *Academic engaged time and its relationship to learning: A review of the literature* (Monograph No. 17). Minneapolis, MN: University of Minnesota for Research on Learning Disabilities.

Henry, T. (1989, October 9). Christopher Columbus did what? Where? When? *The Register Guard* . Eugene, Oregon.

Hunt, K. (1964). *Grammatical structures written at three grade levels.* (NCTE Research Report No. 3). Urbana, Illinois: National Council of Teachers of English.

Johnson, M., Kress, R., & Pikulski, J. (1987). *Informal reading inventories*-(4th ed.). Newark, New Jersey: International Reading Association.

Martorella, P. (1972). *Concept Learning: Designs for instruction.* Scranton, Pennsylvania: Intext Educational Publishers.

Miller, H. G., & Williams, R. G. (1973). Constructing higher level multiple choice questions covering factual content. *Educational Technology, 13*(5), 39-42.

Miller, H. G., Williams, R. G., & Haladyna, T. M. (1978). *Beyond facts: Objective ways to measure thinking.* Englewood Cliffs, New Jersey: Educational Technology Publications.

National Geographic (1989). Go Team Go! *World, 172,* 3-7. Washington, D. C: National Geographic Society.,

Roid, G. H., & Haladyna, T. M. (1982). *A technology of test item writing.* New York: Academic Press.

Seddon, G. M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research, 48*(2), 303-323.

Shapiro, M. (1983). *Basic tips on the American College testing Program: ACT.* Woodbury, New York: Barron's Educational Series, Inc.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction (pp. 169-198).* Baltimore: Johns Hopkins University Press.

Tindal, G., & Marston, D. (1990). *Curriculum-based assessment: Evaluating instructional outcomes.* Columbus, Ohio: Charles Merrill.

Williams, R. G. (1977). A behavior typology of educational objectives for the cognitive domain. *Educational Technology, 17*(6), 39-46.

Williams, R. G., & Haladyna, T. M. (1982). Logical operations for generating questions (LOGIQ): A typology for higher level test items. In G. H. Roid, & T. M. Haladyna (Eds.), *A technology for test-item writing* (pp. 161-186). New York: Academic Press.


## Key Vocabulary

***Assessment***– Systematic collection of information to improve instruction. This information can come from three different sources: Interactive observations, analysis of permanent products, and tests and measures.

***Basic Skills***– Motoric behaviors in academic domains needed to adequately learn material. Four areas are covered: basic reading, spelling, writing, and math skills.

***Content Knowledge***– Information expressed verbally (orally or in writing). Three types of content information are presented: facts, concepts, and principles; six behavioral acts are considered in exhibiting this information (reiterating, summarizing, illustrating, predicting, evaluating, and applying).

***Interactive Observations***– Information collected during instruction through questions and observations.

***Learning Rates***– Three components that define (a) the type of performance outcomes, (b) the amount of these outcomes that has been learned, and (c) the amount of time taken to learn them.

***Mastery***– Performance requirements that reflect adequate (proficient) understanding or manipulation of behavior (skills, knowledge, or procedures).

***Performance Objectives/Outcomes***– The content or activity that forms the focus of teaching and assessment of student behavior. Three different outcomes are included: Basic skills, content knowledge (what students know), and procedural knowledge (how students perform).

***Permanent Product Analysis***– Judgments of quantity and quality through student creations, projects, compositions, etc.

***Procedural Knowledge***– Accurate completion of steps in sequence, which may involve behavioral acts (basic skills) or manipulation of information (content knowledge).

***Tests and Measures***– Structured presentation of items in which responses are scored in terms of correctness or quality. This definition can include both objective and subjective evaluations of performance.

*University of Oregon*