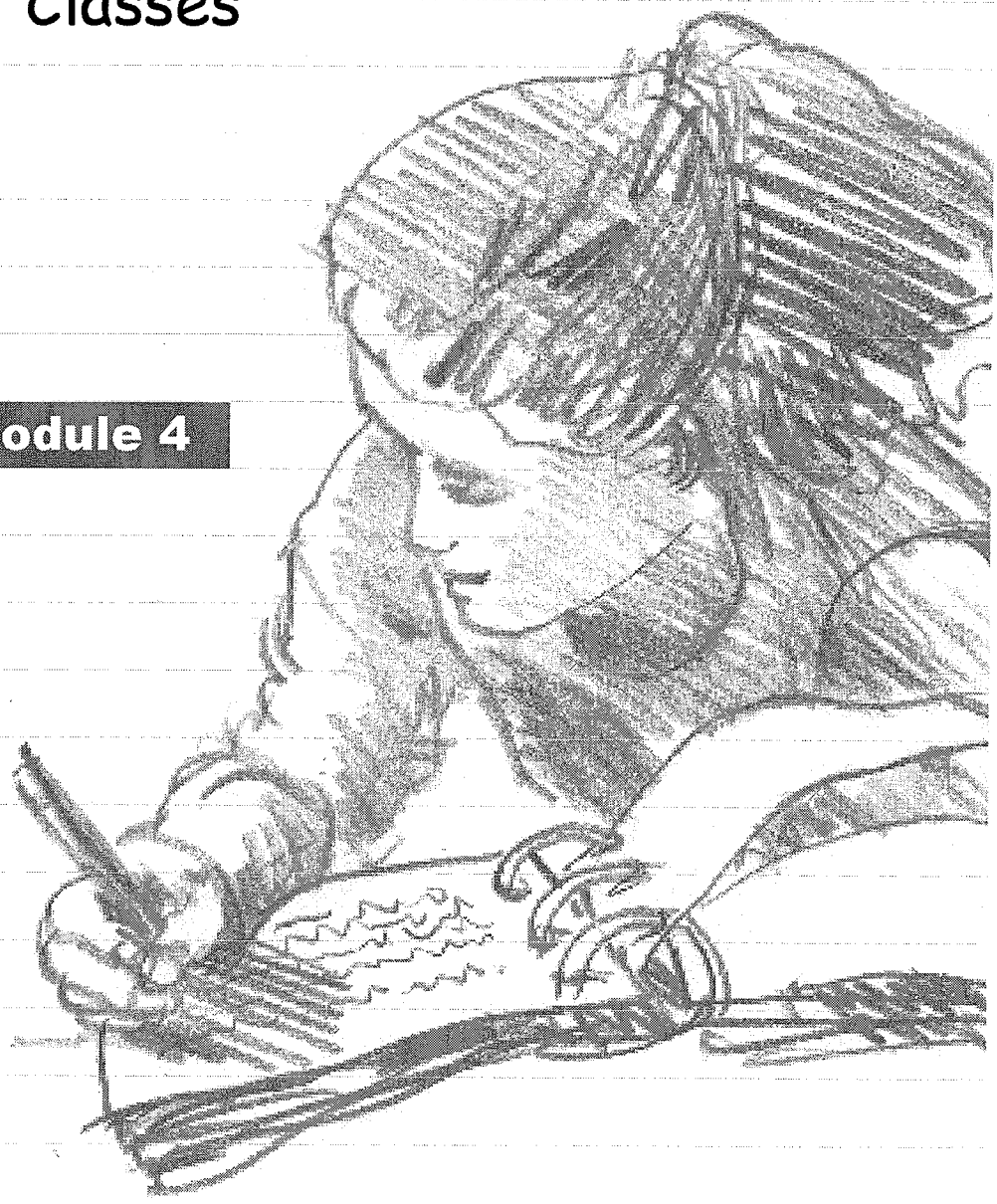
BRT Behavioral Research & Teaching

Focus on Assessment and Learning in Content Classes



Training Module 4

SECOND EDITION

Victor Nolet

Gerald Tindal

Geneva Blake

College of Education-University of Oregon

Published by
Behavioral Research and Teaching
College of Education
University of Oregon

Tindal, G., Nolet, V., Blake, G.
Focus on Assessment and Learning in Content Classes
Training Module 4

Preparation of this document was supported in part by the U.S. Department of Education, grant number H029K10130. 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.

Layout: Geneva Blake and Jerry Marr
Cover design: Barry Geller

Table of Contents

Goals of the Training Module

In *Training Module 3, Focus on Teaching and Learning in Content Classes,*[1] you had an opportunity to adapt existing curriculum materials in ways that allow instruction to focus on complex knowledge forms and higher-level intellectual operations. As part of adapting the curriculum, it was necessary for you to apply your knowledge of the content to specify what it was that you wanted your students to learn.

Eventually, you must ask, "Did the students learn what I wanted them to learn?" Usually, you answer this question by giving your students a test. As you found when looking at curriculum materials in the previous module, the tests included with most textbooks tend to focus on facts and, therefore, do not match the objectives of much of the instruction you give your students. In the absence of other alternatives, teachers find they must rely on their own judgments and observations of student performance.

The present training module picks up where *Training Module 3* left off, offering some concrete guidelines that you can use in developing a variety of assessment items. These items will reflect more closely the content and structure of the instruction you give your students, allowing you to answer precisely the question, "Did the students learn what I wanted them to learn?" The objectives of this training module are summarized below. During the training you will have opportunities to learn about, and gain practical experience with, each of these objectives:

- *Decide what information you need to collect in order to evaluate the effects of your instruction.*
- *Design assessment tasks that provide students opportunities to use intellectual operations beyond reiteration.*
- *Design assessment tasks that are reliable and valid.*
- *Employ quantitative as well as qualitative scoring methods to evaluate student performance.*

---

[1] Tindal, G., Nolet, V. W., & Blake, G. (1992). *Focus on Teaching and Learning in Content Classes* (Training Module No. 3). Eugene, OR: University of Oregon, Research, Consultation, and Teaching Program.

Research Basis for the Training

Serving Special Education Students in Content Area Classes

In a review of literature pertaining to teacher competence, Reynolds[2] identified three domains of understanding associated with teacher competence: (a) broad background knowledge and content understanding, (b) general principles of teaching and learning, and (c) content-specific pedagogy. In current models of special education in middle and high school content classes, expertise in these three domains often has been delegated to various individuals, including assessment specialists, special education teachers, and classroom teachers. Among these domains, classroom teachers are most highly trained in content knowledge and pedagogy, while special educators are likely to be most highly trained in principles of teaching and learning. Indeed, expertise in this area defines the job of special education teachers in many settings.[3] However, special education teachers often assume responsibility for much of the content instruction that special education students receive. For example, in a survey of all state departments of education, McKenzie found that almost half of the schools in the United States use a content approach in their special education programs, and fully 20% of students with learning disabilities received *all* content instruction from special education teachers.[4] In earlier research, Patton, Palloway, and Cronin surveyed 284 special education teachers from all grade levels and found that between 50% and 70% taught social studies in special education settings, but 43% indicated they had no training in social studies education.[5]

---

[2] Reynolds, A. (1992). What is competent beginning teaching: A review of the literature. *Review of Educational Research, 62,* 1-36.

[3] Pugach, M. (1987). The national reports and special education: Implications for teacher preparation, *Exceptional Children, 53,* 308-314.

[4] McKenzie, R. G. (1991). Content area instruction delivered by secondary learning disabilities teachers: a national survey. *Learning Disability Quarterly , 14,* 115-122.

[5] Patton, J. R., Palloway, E. A., & Cronin, M. E. (1987). Social studies instruction for handicapped students: A review of current practices. *The Social Studies Journal,* May/June, 131-135.

As these findings suggest, special education teachers may have limited background in content disciplines such as English, science, geography, or history, and few would be expected to have understandings comparable to that of content area teachers in *multiple* disciplines. Yet content understanding is the primary factor governing student achievement in content classes.[6] At best, typical special education teachers can be expected to provide adequate content instruction in only one or two of the subject areas typically offered in a middle or high school. Therefore, in many content areas, special education students probably will not receive instruction comparable to that of their peers in mainstream content area classes if they receive it from a special education teacher.

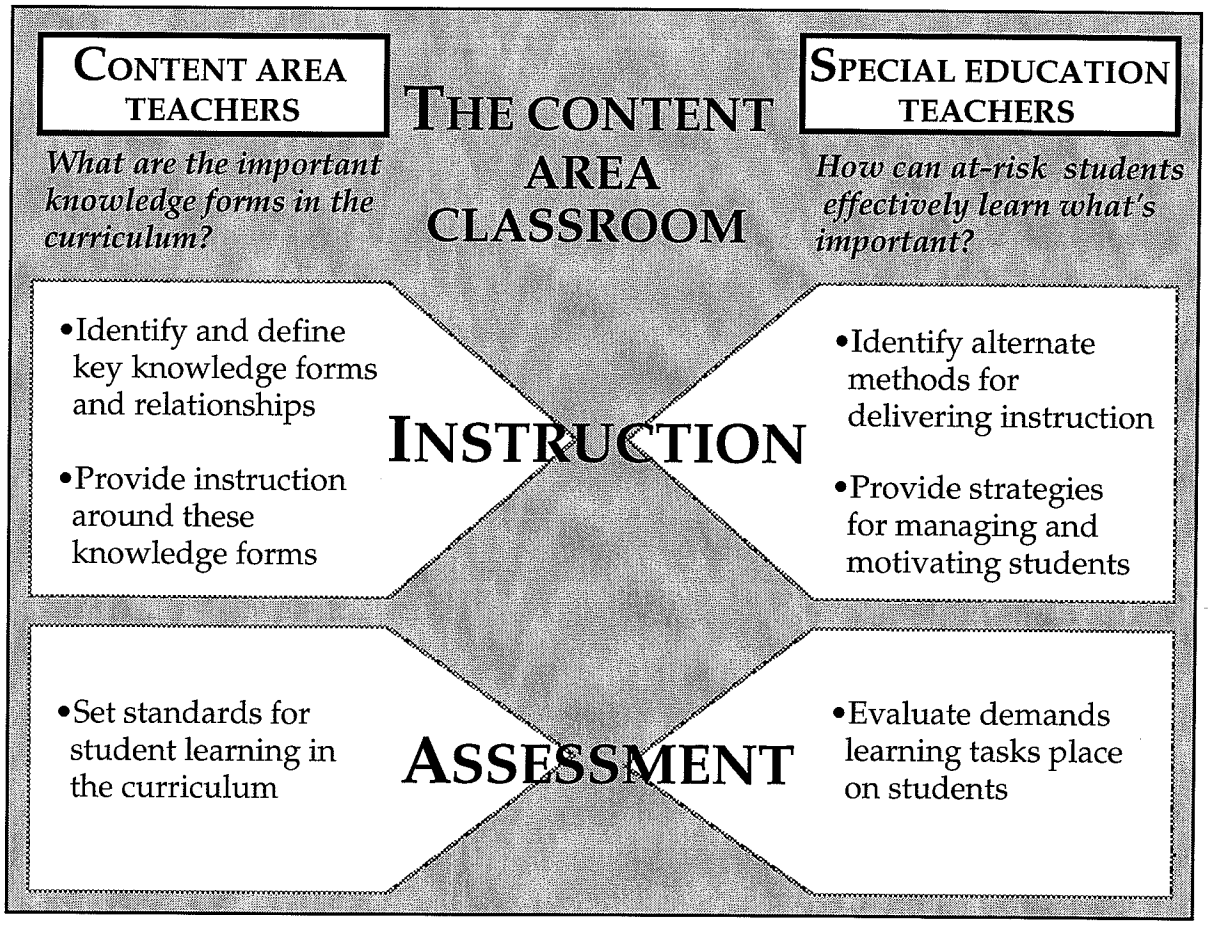
The clear implication is that while special educators may be most effective in implementing strategies that improve the basic reading, writing, communication and social skills of low achieving students, they may be much less effective at helping students apply those skills at a functional level to specific content. William Shakespeare reminds us in Hamlet that "the readiness is all." But, eventually students must shift from "getting ready" to "doing." As many special education students reach middle school and beyond, the focus of their educational program must shift increasingly toward acquisition and use of content knowledge. And it seems appropriate that they should receive instruction from those teachers who have the most expertise in the content areas. However, just as we should not expect special education teachers to have the expertise required to teach in all the content areas, neither should we expect content area teachers to be able automatically to tailor instruction to fit the needs of every special education student who comes into their class.

We propose a model of special education in which content helps to determine, theoretically and conceptually, the purpose of curriculum and instruction. This model defines a new relationship between special education and general education teachers. The content-area teacher brings expertise associated with content knowledge of the particular domain. This discipline knowledge permits identification of key knowledge forms (facts, concepts, principles, and procedures) around which content instruction can be organized. The special education teacher brings pedagogical expertise related to methods for designing instruction, classroom management, and motivational strategies effective with at-risk learners. Both teachers need to consider assessment in terms of the standards for learning (from the

---

[6] Reynolds, *op. cit.*, 1-36.

content teacher's perspective) and the demands made upon students (from the special education teacher's perspective). This model redefines the working relationship between special and general educators. While it is assumed that students receive their primary instruction from the content experts, special educators have a role in helping support students through the curriculum and instruction. They also must become familiar with the demands set by the content area teachers in terms not only of the criterion for success (i.e., performance on the test) but also the content of the assessments. The figure below illustrates this new relationship.



A model of special education for content area classes.

Evaluating the Effects of Instruction

Classroom-Based Assessment

As you move further away from the fact-based approach to designing and delivering instruction in the content area, you will encounter a problem with most of the testing materials that accompany the curriculum from which you may have drawn your instruction: These materials become increasingly inadequate to help you evaluate your students' understanding of the material you have taught them. What you need is an approach to assessment that encompasses the variety of concepts and principles forming the content of your instruction, as well as the complexity of the intellectual operations in which you and your students have engaged during instruction. At the same time, this alternative approach must be flexible enough to allow all students, regardless of their reading and writing skills, to demonstrate what they have learned. Classroom-based assessment is one approach that can be employed to meet these needs.

Classroom-based assessment is a broad term that refers to a variety of procedures for gathering information about student performance. It encompasses a number of methods that can be used to sample instruction delivered in content area classes. Indeed, as you will find in this training module, this may be one of the most important features of classroom-based assessment. Furthermore, the classroom-based assessment procedures you will learn here can be used in both individually referenced as well as norm-referenced evaluation. Thus, classroom-based assessment can be used both to evaluate an individual student's progress over time and compare her performance to that of her peers at a given point in time.

On the following page is a summary of the most important features of Classroom-Based Assessment. The significance of each of these features may not be clear to you at this stage of the training. In fact, the difference between the classroom-based approach and more traditional approaches may not be obvious to you after reading this summary. It is certainly true that traditional assessment *may* include some of these features some of the time. The important distinction of classroom-based assessment is that it must, by definition, include *all* of these features *all* of the time.

Important Features of Classroom -Based Assessment

1. **It samples instruction representatively.**
   This means that the tasks used in classroom-based assessment are a fair sample of the goals of instruction. It implies that classroom-based assessment tests what students are taught.

2. **It is technically adequate.**
   This means it is *reliable* and *valid.* An assessment task that is designed and administered in a reliable manner is relatively free of potential sources of error that have nothing to do with the purpose of the task. A valid assessment task can be used to answer the question: "Did the students learn what I wanted them to learn?" Reliability and validity will be covered in more detail later in this module.

3. **It employs production responses.**
   Students are expected to generate a product as a result of the assessment process. This product could be as simple as a few phrases or sentences or as elaborate as an essay. Production responses, also may include spoken responses, such as may be elicited in a structured interview, as well as nonverbal constructions, such as maps, graphs, and drawings.

4. **It can provide information for making instructional decisions.**
   The information obtained from classroom-based assessment can be used to evaluate the effectiveness of past instruction and to plan future instruction. Classroom-based assessment may or may not be useful for making other decisions, which may be social or political rather than educational (such as assigning grades, or placing a student in special education).

5. **It can be used with a range of evaluation standards.**
   This means that classroom-based assessment can be used to (a) compare an individual's or group's performance to that of a comparison group (norm-referenced evaluation), (b) estimate the extent to which content or skills have been mastered (criterion-referenced evaluation), or (c)chart an individual student's progress over time (individual-referenced evaluation).

In this training module, we will focus primarily on the first three aspects of classroom-based assessment: technical adequacy, sampling plans, and production responses. The last two aspects of classroom-based assessment—making instructional decisions and evaluation standards—will be detailed in a later training module.

Assessment Vocabulary

As you may have noticed already, talking about classroom-based assessment involves a fairly specialized vocabulary. In the preceding short description of classroom based assessment, you were exposed to the terms "assessment," "measurement," "evaluation," "individually referenced," "norm referenced," and "technically adequate." Unfortunately, these terms often are misunderstood and used incorrectly. The first half of this training module is aimed at clarifying the meaning of some of these terms. In the second half of the module, you will have the opportunity to apply some of this new knowledge in designing and scoring various examples of classroom-based assessment.

Let's start with four terms that may be the most misunderstood of all: testing, evaluation, measurement, and assessment. Often these terms are used interchangeably, but each has a specific meaning.

Assessment

Assessment is a general process that involves use of testing, measurement, and evaluation. The Standards for Teacher Competence in Educational Assessment of Students[7] define assessment as:

> the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weaknesses, to judge instructional effectiveness, and curricular adequacy, and to inform policy.

The most important aspect of this definition is the idea that assessment is a process that may involve collection of more than one kind of data, and takes place over time, rather than on a one-shot basis. The second key aspect of this definition is that an assessment process should support a decision-making process. For example, "judging instructional effectiveness" probably requires weighing a number of variables that may include the goals of instruction, time allotted for instruction, amount of information presented, student skills, and the adequacy of resources available. No single score can provide this information. Assessment is a purposeful endeavor aimed at informing educational decisions.

---

[7] American Federation of Teachers, National Council on Measurement in Education, and National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington DC: author.

Testing

Testing is a process of sampling behaviors systematically in an artificial situation intended to reflect the real world. It is not a direct observation of performance.[8] Most definitions of testing emphasize a systematic and structured plan in which narrowly defined items or tasks are presented to a test-taker for the purpose of making inferences about the performance of the test-taker in a larger context. For example, the National Teachers' Exam (NTE) samples behaviors that are thought to reflect competence in teaching. It is not a direct observation of teaching performance.

Tests are intended to provide information that can be used to make some determination or decision, but testing does not in and of itself refer to a decision-making process. Furthermore, testing is not an instructional procedure. *Tests cannot teach.* However, when used to provide *feedback* to the test-taker about performance, tests can provide data for very powerful instruction.

The important feature of testing that distinguishes it from other procedures discussed here is the specificity of focus. The critical variables that must be considered are (a) specification of a domain to be sampled, (b) systematic procedures for selecting tasks to represent that domain, and (c) the presentation of these tasks to a test taker. (The concept "domain" is central to this module. We will devote an entire section to explaining it, and we will provide you with examples and exercises that give direct experience with domains.) Testing must be defined narrowly, to refer to development and administration of a single instrument rather than a larger process of collecting information. However, testing can refer to a range of devices and procedures, including published instruments, end-of-unit tests, and teacher-made quizzes.

Measurement

Measurement is the process of assigning a value to behaviors in such a way as to represent quantity or quality.[9] Measurement always concerns numbers and always concerns *how much* of an attribute is present. Measurement in educational assessment is analogous to other forms of measurement. Suppose you were told that the length of a piece of rope is "30." You would immediately want to know what

---

8

[8] Tindal, G. R. & Marston, D. (1990). *Classroom-based assessment: Evaluating instructional outcomes.* Columbus, OH: Charles Merrill.

[9] Nunally, J., (1978). *Psychometric Theory.* New York: McGraw-Hill.

scale is represented by "30." Obviously, a 30-centimeter rope is much shorter than a 30-yard rope.

Although tests almost always result in a numeric score (such as percent of items correct), this does not automatically mean that testing is synonymous with measurement. For example, a test could be scored "pass/no pass," or with scores of "A," "B," "C," "D," and "F." At the same time, the assignment of a numeric score does not necessarily imply that a measurement process was involved. For example, social security numbers, telephone numbers, and license plate numbers could all be considered forms of scores, but they tell us little more about the person to whom they are assigned than the general geographic area in which they live.

The numbers assigned to a behavior in educational measurement must represent an underlying scale or set of attributes. We assume that the behavior of interest exists on some continuum.[10] For example, a score of "30" on a reading test by itself tells us almost nothing about the reading ability of the person who obtained this score. We need more information about the test, the person, and the definition of reading involved. The score "30" could represent the number of words read, sentences read, books read, minutes spent reading a passage, or number of items correct. In other words, we need to know the scale underlying the score "30."

Consider this example:

> Mr. Jones has four sections of the ninth grade creative writing class. Each week, he administers a writing test in which students write for 10 minutes in response to a story starter. Writing samples are assigned holistic scores of 1, 2, 3, 4, 5, or 6, with 1 the lowest score possible and 6 the highest score possible. Because he has so many papers to score, Jones has asked his assistant, Mr. Smith, to help him score tests.
>
> Unfortunately, Mr. Jones forgot to tell his assistant to rate the papers on the dimension of "creativity" and to ignore spelling and punctuation errors. Smith is a stickler for accuracy and ends up assigning lower scores to the papers with more spelling errors.

All papers are assigned a score ranging from 1 to 6, but the scores assigned by Mr. Jones represent the amount of creativity present while the scores assigned by Mr. Smith represent the amount of accuracy present.

Often we may be interested in assigning a numeric value to the *quality* of a student's performance. For example, we may rate student essays on a scale of 1 to 5 on the dimension of "plot development." Essays in which a well-developed plot is

---

[10] Glaser, R. (1963). Instructional technology and measurement of learning outcomes. *American Psychologist, 18*(3), 519-521.

evident would receive a score of 5, and those in which no plot is evident would receive a score of 1. This process is also a form of measurement. The underlying scale is related to the primary trait, "plot development." Students who have learned the skill of creating a plot in an essay would presumably demonstrate this by writing essays which receive a higher score. Students who have not learned this skill would write essays that receive lower scores.

Evaluation

Evaluation is the process of making a decision, or reaching a conclusion about student performance, based on data obtained from an assessment, testing, or measurement process. Where assessment is the process of *collecting* data, evaluation is the process of *using* data. In this respect, students are not evaluated, rather data about their performance (scores, results, ratings, etc.) are evaluated.

Educational evaluation entails use of criteria or standards by which a particular score or set of scores is compared. Three types of reference standards can be involved. In norm-referenced evaluation, the test scores obtained by specific individual or group of individuals are compared with all others who take the test, under similar conditions. In criterion-referenced evaluation, performance is judged according to a set of standards that represent "competency" or "mastery" in a domain. These standards are set beforehand according to expert judgment. In individual-referenced evaluation, the performance of a particular individual at a specific point in time is compared with the individual's previous performance on similar tasks, performed under similar conditions.

◆◆◆◆◆

Let's return to the first three characteristics of classroom-based assessment listed earlier. We can now be a little more specific in our description. First, because we are talking about *assessment*, we know that the emphasis will be on collection of data that can be used to support educational decision-making. It is likely that we will be interested in multiple forms of data collected at different points in time. Because it is classroom-based, it focuses on what and how students actually are taught.

The reason tasks that require production responses are used extensively in classroom-based assessment is that they provide much more information about students' thinking than selection responses. In particular, classroom-based assessment tends to make less use of tests that employ *indirect* tasks and make more

use of *direct* tasks. The distinction between direct and indirect tasks can be most easily grasped through a couple of examples:

> Suppose you want to decide how well a student has mastered the mechanics of writing. You could give the student a set of five written sentences containing various errors in spelling, grammar, and punctuation. The student's task would be to edit the sentences, correcting the errors. This is an example of an *indirect* task. Alternatively, you could provide instructions (verbal or written) about the kinds of sentences you want the student to write, and then have the student make up a paragraph that contains the desired characteristics. This is an example of a *direct* task.

> Here is another example. Suppose you want to find out how well students can use information you provided during instruction about the effects of water pollution. You could test this *indirectly* by asking them to select the correct response(s) from a list of possibilities. Or, you could construct a novel situation involving water pollution and ask students to make inferences about the possible effects. This second task is more *direct*, because it requires students to produce a response that uses content knowledge you provided during instruction.

In later sections of this module, you will have many opportunities to see examples of production tasks, and gain experience in designing tasks that you can use in your classroom to assess student learning.

The goal of classroom-based assessment is to link measures of performance more closely to the instruction students receive in the content classroom. It seems reasonable to assume that, by asking students to display directly the behaviors we want to measure, we will be able to gauge more accurately whether they have learned what we wanted them to learn.

But, what is it that we wanted them to learn? And how can we be sure that the assessment tasks we design and administer are reliable and valid gauges of students' learning? In the next two sections, we will present a framework for defining what is taught in content classes. We will build on the use of knowledge forms and intellectual operations that were developed in *Training Module 3* as guiding principles for planning and designing instruction. Not surprisingly, many of the aspects of planning for and designing instruction can be extended to planning for and designing assessment. In fact, if you have gone through the steps of instructional planning and design carefully, much of the work of developing assessment tasks is already done for you. Instruction links to assessment by identifying an *assessment domain* and a *plan for sampling* from that domain. These two concepts are central to classroom-based assessment and are the topics of the next two sections.

Defining the Domain in Content Classes

The primary goal of classroom-based assessment is to provide information to systematically answer the question, "What have my students learned?" Usually, we are not interested in what students have learned about "life, the universe, and everything"[11] merely as a result of being alive and awake. We want to know specifically what students have learned as a result of attending certain classes in school.

One way we add specificity to the question of what students have learned is with reference to a fixed period of time. We may be interested in finding out what students have learned during a relatively short period of time, for example during the last two weeks, or we may want to know what students have learned during the last six months. Obviously, one of the goals of education is to move students along a continuum of learning over time. We assume that students enter class "not knowing" and, if instruction has been effective, leave the class "knowing." Schools are structured for this kind of time-based thinking, with events such as mid-term exams and quarterly grades. (We will have more to say about the use of grades when we examine reliability and validity in later sections.) However, linking assessment to a specific period of time is of limited utility. References to time simply represent a convenient notation for describing instruction.

We really are interested in a more complex version of the question, "what have my students learned?" Generally, we want to make reference to a specific body of information or level of performance. Underlying the issue of what students have learned during the last two weeks, lurk questions like these:

- What have my students learned about biomes?
- Did my students learn the most important information about the Middle Ages?
- Have my students learned to write more persuasive essays?

A more specific goal of assessment, then, is to answer the question, "Did my students learn the things I wanted them to learn?" Again, we make the assumption that

---

[11]Adams, D., (1983). *The hitch-hiker's guide to the galaxy.* New York: Harmony Books.

students enter class not knowing and, if instruction has been effective, they leave some time later knowing *about something* or *how to do something*.

Because any two students enter a learning situation with different skills and background knowledge, and because the effectiveness of instruction may differ from teacher to teacher, curriculum to curriculum, and student to student, we rarely view learning as a dichotomous "either-or" situation. We expect students to show different levels of performance after instruction than they did before instruction; and even after instruction, we expect to see different students exhibit different levels of competence. In other words, we view knowing *about something* or *how to do something* as existing on a continuum of performance.

While grades may communicate the notion of a continuum of performance, they aren't very useful for answering questions such as "Did my students learn the most important information about the Middle Ages?" To answer this question, we need to specify two things: First, we have to clarify what we mean by "the most important information about the middle ages," and, second, we have to specify what we mean by "learn." Such a level of specificity requires clear delineation of the domain in which assessment is to occur.

Domains in Instruction and Assessment

In the world of measurement and assessment, the term *domain* has a specific meaning, but it is not easily summarized. Usually, *domain* refers to an *assessment domain.* However, as you will see, we also can talk about an *instructional domain.* We'll explore the meaning of both of these terms. Let's start with assessment domain. Here is how we define the term *assessment domain:*

> An *assessment domain* is a specific set of skills or body of knowledge associated with an instructional intervention. This domain also may include a continuum of competence in using the skills or knowledge.

Our definition of *assessment domain* has two components that require discussion. First, an assessment domain should be linked to a specific instructional intervention. For example, an assessment domain could pertain to "Chapter 13 in the social studies book," or "three months of instruction on writing compare-contrast essays," or "seven lessons on fossil fuels and alternative energy." In some cases, an assessment domain could involve instruction delivered during an entire school year (for example, "reading instruction delivered in the third grade at the King

Elementary School"). This instructional linkage differentiates assessment domains from broader conceptions of skills and knowledge delivered in schools. For example, we probably would **not** classify the following as domains in and of themselves: "reading," "math computation," and "penmanship"—although there may be times when the domain could be defined as broadly as these terms would suggest.

Second, our definition of domains involves the *use* of skills or information as well as the *information* itself. Here is an example that illustrates this point:

> You are a music teacher and you have been working with a group of beginning piano players. You might evaluate the success of your instruction by asking each student to perform in front of the class for five minutes at least once each week. Each student would perform the same piece of music, selected by you, under the same "test" conditions.
>
> The assessment domain could be defined in an number of ways. On one hand, you could define the domain according to the music the students would be expected to play. The domain could be described as simply as consisting of those passages students have practiced recently (for example all passages on pages 5-12 of the student's music book).
>
> Alternatively, you could define the domain according the content of what you have taught but attend only to the characteristics of the music (for example, music on the G clef; pieces that contain quarter, half, and whole notes; or pieces in the key of C).
>
> Finally, because piano playing generally is judged according to the skill of the piano player, as well as the difficulty of the music, you might want to evaluate your instruction on the basis of students' technical skills at striking the piano keys, maintaining control of rhythm, and positioning the fingers of the left hand.
>
> To account for each of these aspects of your instruction, you would have to define the assessment domain according to some continuum of competence in playing the piano as well as according to the characteristics of the music the students play.

Knowledge Forms and Intellectual Operations Revisited

Students' academic performance in content classes can be analyzed for the use of skills or information, as well at the information itself. In *Training Module 3*, we said that content knowledge consists of facts, concepts, and principles, and that students could demonstrate use of these knowledge forms in six intellectual operations: reiteration, summarization, illustration, prediction, evaluation, and application. A brief summary of each of the intellectual operations is provided in the table below. In addition, we presented a two dimensional grid (reproduced on page 16) that shows the interactions between intellectual operations and knowledge forms. In *Training Module 3* this grid served as a guide for analyzing the curriculum and planning instruction, i.e., defining the instructional domain. It also can be used to define the assessment domain in content classes.

Intellectual Operation	refers to the behavior employed in using or manipulating knowledge forms.
Reiteration	*Verbatim reproduction of material that was previously taught.* • The emphasis is on *verbatim*. The wording in the student's response must be very nearly identical to that presented in instruction.
Summarization	*Generation or identification of a paraphrase, re-wording, or condensation of content presented during instruction.* • The emphasis here is on previous presentation of material. Therefore, summarization involves remembering information to a much greater extent than manipulating it.
Illustration	*Generation or identification of a previously unused example of a concept or principle.* • The emphasis here is on use of an example that was not presented in instruction. In this respect, the student is expected to employ information about the attributes of a particular concept or principle rather than to simply remember whether or not an event exemplifies a knowledge form.
Prediction	*Description or selection of a likely outcome, given a set of antecedent circumstances or conditions that has not previously been encountered.* • Again, the emphasis is on the *use* of information in a novel context rather than remembering a response from previous instruction.
Evaluation	*Careful analysis of a problem to identify and use appropriate criteria to make a decision in situations that require a judgment.* • Evaluation focuses on decision making. The student first must recognize or generate the options available and then use a set of criteria to choose among them.
Application	*Description of the antecedent circumstances or conditions that would be necessary to bring about a given outcome.* • Application is the reverse of prediction. The student must use information about a concept or principle to work backwards from the circumstances presented and tell what happened to create it.

Intellectual Operations	Knowledge Forms: Facts	Concepts	Principles
Reiteration	Yes	Yes	Yes
Summarization	Yes	Yes	Yes
Illustration	No	Yes	Yes
Prediction	No	Yes	Yes
Evaluation	No	(Yes)	Yes
Application	No	(Yes)	Yes

To see how this grid might be used to define an assessment domain, consider the following example:

> A tenth grade social studies teacher defines the instructional domain for a unit on the history and geography of northern Africa as consisting of 10 key facts, 8 concepts, and 6 principles, based on that teacher's perception of the information most vital for understanding why people live in certain areas of Sub-Saharan Africa. The domain also includes a number of intellectual operations associated with each knowledge form. At the end of the unit, the teacher expects students to be able to manipulate the various facts, concepts, and principles using summarization, illustration, prediction, evaluation, and application.

An outline of what the domain might look like is given on the next two pages.

Key Facts about Northern Africa	*Intellectual Operation*
1. The Sahara is the worlds largest desert.	Summarize
2. Dry lands are usually found on the western coasts of continents between 20° and 30° north and south of the equator.	Reiterate
3. The Sahara receives 4 to 8 inches of rain (or less) each year.	Reiterate
4. Sahel is an Arabic word that means "plain" or "shore."	Summarize
5. The annual rainfall in the Sahel varies from 4 to 23 inches per year.	Summarize
6. Erosion caused by overgrazing is one of the most important ecological problems facing the Sahel.	Summarize
7. Soils in most tropical areas are not very fertile because of constant leaching from rains.	Summarize
8. A shortage of firewood in the Sahel has caused depletion of the acacia trees.	Summarize
9. The process of over-grazing or cultivation of marginal land, and, subsequently, the land reverts to desert is called *desertification.*	Reiterate
10. The people who live in the Sahel are mostly nomads.	Summarize

Key Principles about Northern Africa	*Intellectual Operation*
1. If the rainfall in an area is extreme (very little or very much), the soil will contain fewer plant nutrients.	Illustrate
2. The more plant nutrients the soil in an area contains, the more favorable the area will be for vegetation (diversity, value).	Predict Apply
3. The more extreme the climate in an area, the less favorable the area will be for vegetation.	Evaluate
4. If there is great variation in relief in an area, the population will be less dense.	Illustrate
5. If the vegetation in an area is valuable (provides food or shelter), the population will be more dense.	Predict
6. If the climate in an area is extreme, the population will be less dense.	Illustrate Predict

Key Concepts Related to Africa				
Concept	*Attributes*	*Examples*	*Non Example*	*Intellectual Operations*
Relief	• Differences between high and low places on the earth.	• Alps Mountains • escarpments • Nile River	• Belgium • Africa • Cairo	Summarize Evaluate
Climate	• The combination of heat energy and moisture that a place receives over time. • Climate involves the seasonal distribution of temperature and precipitation as well as annual values. • Patterns of climate can be classified with standard categories such as Marine, Tropical, Continental, etc.	• Humid continental climate regions experience extreme temperature fluctuations and receive more moisture in summer than in winter. • The Sahara receives less than 8 inches of rainfall each year and temperatures range from 10°C to 31°C.	• The average temperature of Tallahassee, Florida during February is 56°F. • Dry lands are usually found along the western coasts of continents between 20° and 30° north and south of the equator.	Summarize Illustrate
Vegetation	• The plants and trees that grow naturally in an area. These are determined by a combination of soils and climate.	• savanna grasses in sub-Saharan Africa • coniferous forests in the American Northwest.	• sugar cane grown on plantations in Zambia • grapes grown in vineyards in California.	Illustrate
Soils	• The surface layer of mineral and organic matter that is the area occupied by plant roots. Poor soils provide few plant nutrients and rich soils provide many plant nutrients.	• The soil found in the savanna contains minerals but does not store plant nutrients well.	• Unweathered rocks contain no organic matter.	Summarize Illustrate Predict
Population Distribution	• Differences in population density from one place to another, within a specific area.	• Most Canadians live within 200 miles of the U.S. border • The population density of the Nile River valley is over 50 people per square km. In the rest of Egypt, the population density is less than 1 person per square km.	• 10.7% of the world's population live in Africa. Africa contains 22 percent of the world's total land mass. • The population density of the United Kingdom is 225 people per square km.	Summarize Illustrate

As you can see from this outline, some intellectual operations make more sense with some knowledge forms than others. For example, facts can only be reiterated or summarized. (Obviously, there are just so many ways to say that the Sahara is the biggest desert in the world.) Some concepts have more explanatory power than others and therefore may be used in more than one intellectual operation. For example, the attributes associated with vegetation in this teacher's domain probably limit the uses students can make of the concept, and the teacher may be satisfied if students can give examples of vegetation in a broad range of contexts. On the other hand, on the basis of the attributes the teacher developed for soil and population distribution, these concepts lend themselves to use in a variety of contexts. Finally, because principles represent the glue that binds concepts and facts, these can be used in a variety of more complex manipulations. A key point here is that *the domain consists of* ***both*** *the knowledge forms* ***and*** *the intellectual operations applied to the knowledge forms.* It consists of the information that you want students to learn *and* the manner in which you want them to *use* that information.

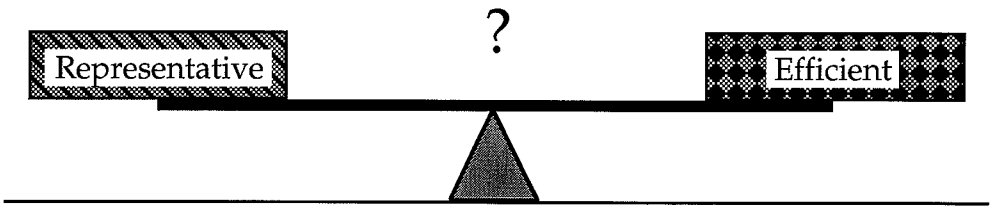
Perhaps the distinction between "instructional domain" and "assessment domain" seems blurred at this point. We introduced the above example as a means of illustrating an "instructional domain." You may have already inferred that the two domains are (or should be) the same. It only makes sense, doesn't it? Where else but from the instructional domain should you draw items for assessment? So, to emphasize a point made earlier, if you have gone through the steps of instructional planning and design carefully, much of the work of developing assessment tasks has been done already for you. When you decide what you want students to learn, you have to decide what you want them to do with the information. When you have identified both the knowledge forms and the intellectual operations, you have defined both the instructional and assessment domains.

Sampling Plans

Linking the Sampling Plan to Instruction

Once you have described the domain instruction and assessment, the next step is to devise a plan for deciding how much of that domain will be included on a test or other assessment task. We have described domains in content classes as consisting of knowledge forms and intellectual operations. We have emphasized that the assessment domain serves as a road map for planning instruction.

When it's time to design assessment tasks, the teacher must decide what aspects of the domain to use. An assessment task merely represents a sample of a broader domain. Usually, the domain contains much more information than the assessment task. The problem confronting the teacher is illustrated in this diagram:



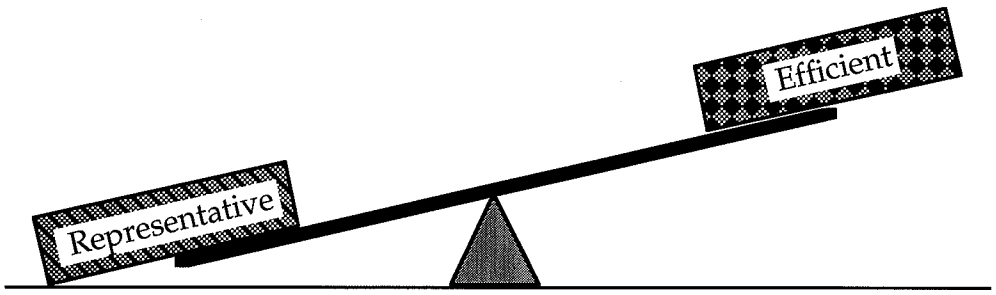
On one hand, the assessment task must be *representative* of the domain. That is, it must reflect the instruction that was delivered, both in terms of knowledge forms and intellectual operations. On the other hand, the assessment task must be *efficient*. It cannot take too much classroom time, and it must be formatted so that the information the teacher gains is worth the effort required to gather it. An effective sampling plan can help you strike a balance between these two demands.

There are three general strategies for sampling a domain to create an assessment task: exhaustive sampling, random sampling, and stratified sampling. We will describe each of these strategies separately. However, these strategies are not necessarily mutually exclusive, and, in fact, some of the most effective sampling plans are made by combining these strategies.

Exhaustive Sampling

Contrary to what you may be thinking, *exhaustive sampling strategy* does not refer to the way you'll feel after you finish reading this training module. The term, *exhaustive,* refers to the amount of the assessment domain included in the assessment task. With this strategy, everything in the domain is chosen. Literally, the domain is sampled exhaustively. If you were interested in sampling exhaustively the social studies domain presented in the previous section, you might construct a test that includes at least 29 items. It would contain at least 10 items that test facts (using summarization or reiteration), 11 items that test concepts (one for each concept, used in each intellectual operation specified), and 8 items that test principles (1 for each principle, used in each intellectual operation specified).

An exhaustive sampling plan assures a very representative sample of the domain. There is a complete overlap with the information the teacher specified as part of the domain and the information that is used in assessment. If the domain is relatively small and relatively well defined, an exhaustive sampling plan may make sense. The problem with this kind of strategy is that, depending on the size of the domain and the format of the items, it could require a lot of time to create an assessment task, a lot of time for students to complete the task, and a lot of time to score. The diagram shows the problem an exhaustive sampling strategy can create. The assessment process is tipped toward representativeness, but at the cost of efficiency.



Random Sampling

One way to get around the problems inherent in an exhaustive sampling plan is to use a smaller sample of items. Of course, as soon as you eliminate some of the domain's information from the assessment task, you create the possibility that the assessment task is not representative of instruction. A *random* sampling strategy

introduces the element of chance to increase the probability that the assessment task maintains a balance between efficiency and representativeness.

In a random-sampling plan, information is chosen for inclusion on the assessment task according to a random process. Suppose you wanted to create a task that has fewer than 29 items but that has a reasonable probability of representatively sampling the social studies domain presented above. You might decide on a 12-item test. Here are the steps you would follow:

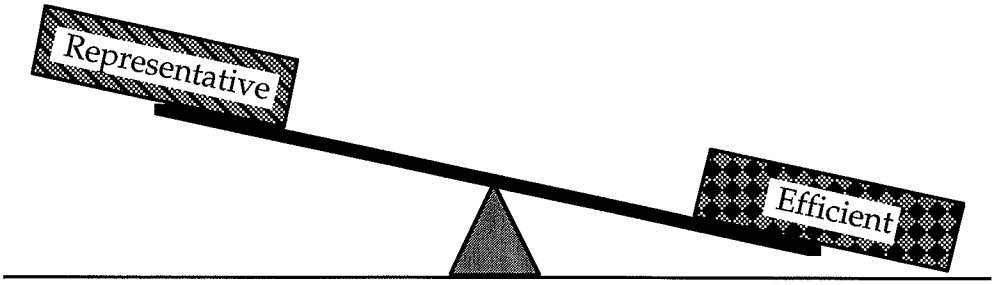
1. Make up a list of all configurations of knowledge forms and intellectual operations (there will be 29 different combinations).

2. Number the items on the list from 1 to 29.

3. Using any random process (for example, a table of random numbers, telephone numbers listed under the "Q's" in the phone book, the last two digits of the license plate numbers of the next 12 cars that pass by your window, etc.), begin choosing items from your list. For example, if your random process gives you the number 5, choose item number 5 from your list of knowledge forms combined with intellectual operations. In the domain presented earlier, this item might require students to "summarize the annual rainfall in the Sahel."

4. Using random numbers to tell you which item to pick, keep choosing items until you have 12.

A random sampling strategy is based on the theory that in any truly random process for choosing numbers, all numbers have an equal probability of being selected. Each time you choose an item from your list, there is an equal probability that *any* item can be chosen. In theory, this strategy allows you to representatively sample a domain without sampling it exhaustively. This is particularly useful if the domain is large. However, because the process is random, by chance alone, you could end up with a test that includes most of the facts but none of the principles, or that requires students to summarize but never asks them to use more complex intellectual operations. In other words, despite your best efforts, the test could still end up different than you might like. Furthermore, a random sampling plan treats everything in the domain as equal. Reiteration of a fact has as great a likelihood of being chosen as a application of a principle. This does not reflect the way most content teachers think about the domains in which they provide instruction. The diagram shows what may happen with a random sampling strategy. The balance is tipped toward efficiency but perhaps at the expense of representativeness.



You might try to avoid this problem by deciding to pick only certain parts of the domain. However, as soon as you "stack the deck" in favor of your view of what the test should look like (for example, by limiting the number of facts you include, or requiring more evaluation than summarization), you may *reduce* the probability that the assessment task will be representative. On the other hand, you may *increase* the representativeness of the task by choosing knowledge forms and intellectual operations that more closely reflect the instruction you provide. The point is, if you are going to intentionally pick some things in the domain, and not pick others, you should use a systematic process for doing so.

Stratified Sampling

A *stratified* sampling plan implies that you have decided that everything in the domain is *not* equal. You impose some structure on the domain that "stratifies" it. Each "stratum," or section, contains things that are similar to one another, but there are differences among strata. In a stratified sampling strategy, you pick things from the domain according to how you want the assessment task to represent each section.

The framework we have presented in this training module and in *Training Module 3* is an example of a way to stratify a domain. By imposing the structure of knowledge forms and intellectual operations on content, instruction and assessment can be systematically organized. One section of the domain contains facts, another contains concepts, and another contains principles. Each of these sections is further stratified by intellectual operations. Using a stratified sampling strategy, you could create an assessment task that includes reiteration of three facts, illustration of five concepts, and application of two principles.

The key to using a stratified sampling strategy is that you make a decision ahead of time about how you will stratify the domain and how you want each stratum, or section, of the domain to be represented in assessment. In practice, this is the way most teachers make up tests. They choose items that they feel are important and they reject items they feel are less important. A potential problem with this

process is that, often, what the teacher *thinks* is representative of instruction does not at all reflect what actually happened in the classroom. For example, we have found in one study that end-of-chapter tests often contain concepts that were never presented in instruction, were not described in the textbook, and weren't even identified by the classroom teacher as important for students to learn.[12] Not surprisingly, fewer than half of the students who took the test correctly answered items related to these concepts.

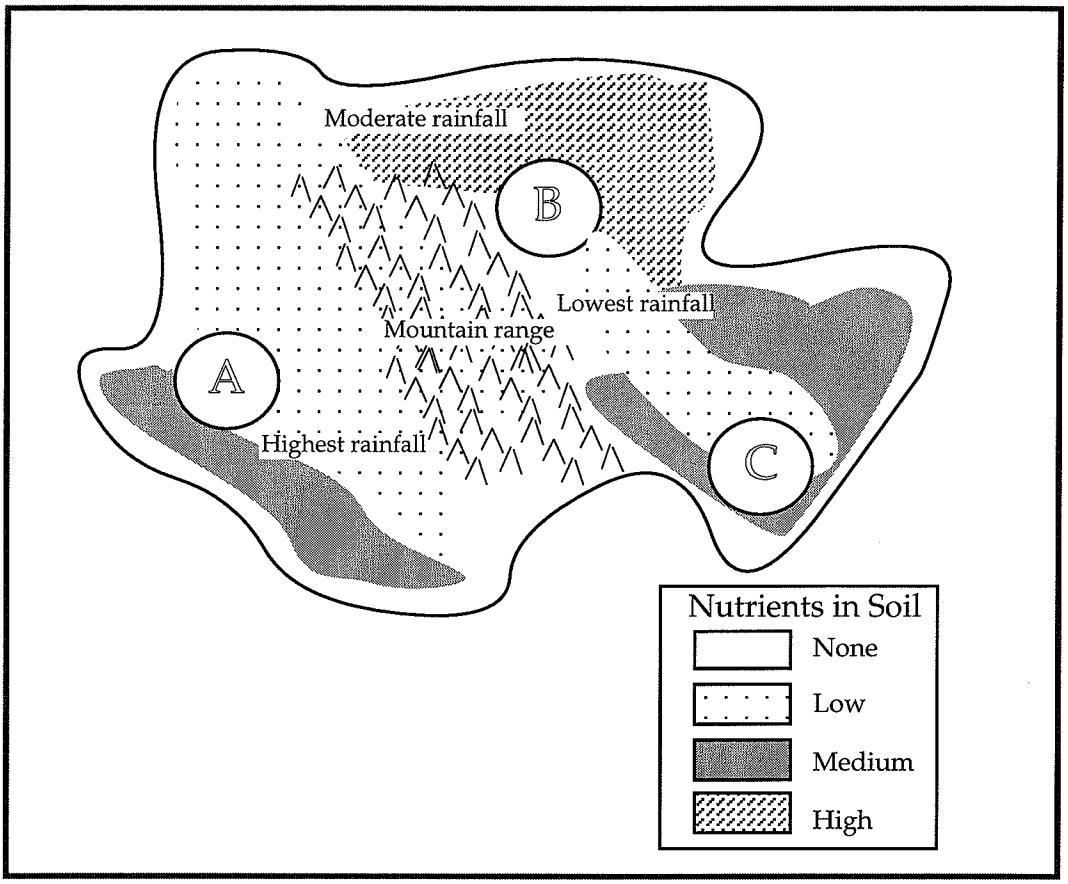
Even if the assessment task samples the knowledge forms representatively, there may still be a problem if the manner in which these knowledge forms are to be used is not also clarified. Consider an assessment task based on the domain outlined in the previous section that includes the item on the next page.

This item requires students to make a prediction based on the principle that the more plant nutrients the soil in an area contains, the more valuable and diverse the vegetation that grows in that area will be. This assessment task involves a knowledge form *and* intellectual operation from the domain. However, this task may not be representative of the domain, depending on how instruction was presented. First, if they have not practiced making predictions with this principle in a *range of contexts*, students probably will be unable present a cogent rationale for their choice of the area with the best vegetation. A large body of research indicates students don't generalize across contexts. If you want students to generalize from a familiar context to a novel one, you need to provide instruction to make it possible. If instruction focused only on the context of soils and vegetation in the Sahel and tropical rain forests of Africa but failed to provide students with a range of non-African examples in which the principle also holds (for example the Mojave desert, and rain forests in Brazil), then an assessment task like the one on the next page may be too novel for students.

Second, students may not be able to respond to the task if the information was formatted as a different kind of knowledge form during instruction. For example, instruction provided in the textbook and by the teacher may simply present the *fact* that, "The vegetation in the Sahel is low in nutrient value because the soil contains few nutrients." Recall from *Training Module 3* that facts can only be reiterated and summarized. Unless students have learned the principle as a lawful relationship (if A then B), they may not be able to use it in complex intellectual operations.

---

[12] Nolet, V. W., & Tindal, G. (1993, February). *Variables associated with student performance in content classes: Description of a research methodology*. Paper presented at the Pacific Coast Research Conference, Redondo Beach, CA.



Here is a soil map for a small island country. In some regions, the soil contains no nutrients. In other regions, the soil is high in nutrient value. Decide where you think the most valuable and diverse vegetation will grow. Do you think it will grow in the area labeled by the letter A, the area labeled by the letter B, or the area labeled by the letter C? Make an X beside the letter of the area where you think the most diverse and valuable vegetation will grow.

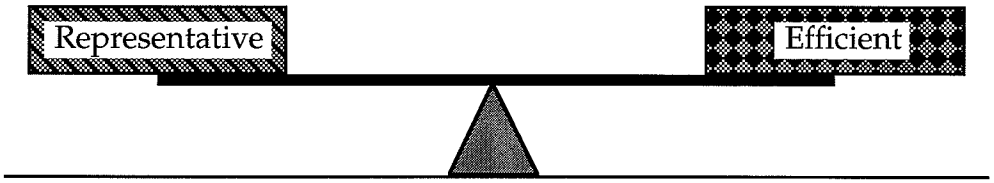
_____A

_____B

_____C

Tell why you made the choice you did. If you said the best vegetation will grow in the area labeled by the letter A, tell why. If you said the best vegetation will grow in the area labeled by the letter B, tell why. If you said the best vegetation will grow in the area labeled by the letter C, tell why.

One way to avoid this kind of problem is to decide *before* instruction is delivered what the domain looks like and how it will be stratified. We presented a content planning worksheet in *Training Module 3* that can facilitate this process. If, during the process of planning for instruction, the teacher identifies the *most important* facts, concepts, and principles for students to learn, there is a greater likelihood that these will be presented in instruction. If the teacher also links these knowledge forms to intellectual operations, then there is a greater possibility that instruction and assessment can be linked systematically. Then a balance can be achieved between representativeness and efficiency. It will be possible to design and use assessment tasks that allow you to determine, with a high degree of confidence, whether your students learned what you wanted them to learn.



Combined Strategies

A stratified sampling strategy is probably the most valuable for creating assessment tasks in content classes. It allows more systematic planning for linking instruction to assessment and can be used with the framework we have presented for analyzing knowledge forms and intellectual operations. However, there may be times when a combination of strategies is useful. In a stratified plus exhaustive strategy, you might choose all the information in one stratum of the domain but use only a sample of items from another part of the domain. For example, you may decide that it is critical for students to be able to summarize and illustrate all of the concepts presented in the domain above, so you would sample these exhaustively for both instruction and assessment. You may decide that the facts listed in the domain outline are of marginal importance and not sample these at all. As an alternative strategy, you may sample one section of the domain randomly (for example, reiteration of facts) but other parts of the domain exhaustively (for example, prediction of principles).

By now the link between instruction and assessment should be clear: the domain of instruction *is* the domain of assessment. Items used to design assessment tasks are drawn from the domain of instruction that you outlined during the

planning process. The strategies described above can be used to construct a plan that *efficiently* samples the domain in a way that is *representative* of the instruction you give your students.

The entire process—from planning for instruction to designing assessment tasks—all boils down to one question: "What's important?" That is, what knowledge do your students need to acquire, what skills do they need to be able to demonstrate in order to understand and make use of content information? When you determine what's important, that can guide you as you plan instruction and assessment.

Validity in Classroom-Based Assessment

An understanding of validity is the key to using classroom-based assessment effectively. Validity underlies all aspects of planning, implementation, and interpretation of assessment. In the context of assessment, validity pertains to the extent to which a procedure measures a domain of interest. Generally, validity asks the question, "Does the test really measure what it is intended to measure?" Applied to evaluating the effects of instruction, we can rephrase this question to ask, "Does this test tell me whether my students learned what I wanted them to learn?"

While validity often is treated as an attribute inherent in *tests*, it is more accurate to refer to the validity of *decisions* based on assessment data than the validity of a specific measure.[13] In this sense, validity is a relative rather than absolute notion: Some inferences are more or less valid than others, depending on the strength of the evidence available to support them. Scores obtained from a particular assessment procedure may be useful (valid) for one type of decision but not for others. A variety of data are examined and an evaluation is made of the extent to which there is a case for the validity of a particular inference.[14] The notion of validity may be the most important development in assessment in recent years. We will present an overview of traditional ideas of validity as well as a more updated view.

Traditional Notions of Validity

We need to be concerned about validity whenever an assessment procedure is being used to measure a quality or construct that cannot be given a clear-cut definition. In educational applications, the term "construct" has referred to relatively intangible conceptualizations of human behavior that cannot be observed directly,

---

[13] Cronbach, L. J., & Meehl, P. E. (1989). Construct validation in psychological tests. *Psychological Bulletin*, 52, 281-302.
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-104). New York: Macmillan.

[14] Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,*, 81-105.

such as "creativity" or "intelligence." Constructs are defined implicitly by a complex network of relationships among a range of variables. For example, suppose you wanted to find out how literate the students in your classroom are. Literacy is an abstract term that could include a wide range of variables including knowledge about culture, reading skill, writing proficiency, reading habits, and use of mature vocabulary.

Because the "literacy" of students cannot be measured directly the way that weight or visual acuity can, it must be inferred indirectly from a variety of types of evidence that we associate with literacy. Such measures could include observable indicators such as measures of reading fluency, or writing competence. You could ask students to fill out a reading questionnaire, on which they provide information about the kinds of books they read and the number of hours a week they read for pleasure. You could give students a cultural literacy quiz on which they summarize key facts that you determine are essential. You could interview each child's primary care giver to find out how much reading and writing takes place in the home. You could ask the librarian how often your students take books out of the library. You could review school records to see how many of your students have been served in Chapter 1 or Special Education for reading or writing difficulties.

Finally, the validity of your inferences about the literacy of the students in your classroom might include your own theory about what constitutes literacy. For example, if the students really are literate, you might hypothesize that they might be more attentive and actively engaged during reading and writing class and earn higher scores on published norm-referenced tests. You could collect data on these variables too.

The validity of your inferences about your students' literacy would be strengthened or weakened by the extent to which the relationships among key variables follow a predictable pattern. You might doubt the veracity of student statements that they like to read if the class has lower than average rates of signing books out of the library. In other words, construct validity is supported when "things that should go together do" and "things that shouldn't go together don't." The entire network of variables must be considered to decide whether your inferences are valid.

Obviously construct validity is no easy thing to establish. Therefore, test developers often look for "predictors" of construct validity. The most important predictor for designers of classroom-based assessment tasks is content validity. *Content validity* refers to how closely an assessment procedure matches the domain it is intended to sample.

We talked about this aspect of assessment earlier, in the section on sampling plans. The more representative an assessment task is of a particular domain, the greater the content validity of inferences based on the results of that measure. For example, when you use an exhaustive sampling plan, you are ensuring a high degree of content validity. Remember, in classroom-based assessment, we are interested in all aspects of instruction, including instruction provided by the teacher and incidental in the course of classroom activities, outside of the textbook.

If your sampling plan is well thought out and linked to instruction, your inferences about the performance of students on an assessment task are strengthened. On the other hand, if your sampling plan does not result in measures that are representative of the domain, the validity of the inferences you make about student performance is reduced. On the next several pages is an example of an assessment task that was used in a 2-week unit on insects during a summer science academy for low-achieving and at-risk sixth and seventh-grade students. Look at the task and think about the following questions:

1. What instruction would you need to provide students for them to be able to complete the task?

2. What kind of sampling plan probably was used to develop this task?

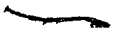
3. How would you determine the content validity of any inferences you might make on the basis of student performance on this task?

Trouble with Insects: You Be the Entomologist

Pretend you are an entomologist and your job is to help people with their insect problems. Recently, people have been having some problems with three insects, but not enough information has been collected yet to be sure which insect is causing each problem. Your job is to use the available information to decide which insect is causing each problem.

So far it has not been possible to get actual samples of the insects, but you do have pictures of body parts from each. Here are pictures of body parts from the three insects. They are labeled Insect A, Insect B, and Insect C. Look at the body parts, and make some notes about your observations. Decide what kind of food each insect eats. Decide how each insect moves around. Decide what kind of habitat each insect lives in.

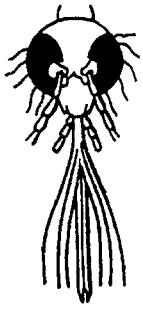
INSECT A

		
Scaly Wings	Siphoning Mouth	Small Legs

INSECT B

		
Parchment Like Wings	Chewing Mouth	Large Rear Legs

INSECT C

		
Membranous Wings	Piercing-Sucking Mouth	Small Legs

Here is the information you have been given about the problems these insects are causing. Decide which insect is causing each problem and tell how you made your decision.

Problem 1: The insect that causes this problem flies and lays eggs in deciduous trees (trees that have broad leaves). When these eggs hatch, the larvae eat the leaves of the trees. They have huge appetites and can strip a tree of all its leaves in a very short time. Sometimes the trees die.

Which insect do you think is causing this problem? (circle one)

Insect A Insect B Insect C

What makes you think so?_____

_____

_____

Problem 2: This insect lays eggs in pools of water. The adult eats blood from living mammals, including humans. One adult of this insect may consume blood from many victims. Sometimes, this insect spreads diseases among its victims.

Which insect do you think is causing this problem? (circle one)

Insect A Insect B Insect C

What makes you think so?_____

_____

_____

Problem 3: This insect lives in tall grasses and escapes predators by hopping. The adult of this insect eats grass, flowers and leaves. Sometimes, when climate conditions are right, there many thousands of this insect may hatch in a very small area. During these times, crops of wheat and grasses may be ruined.

Which insect do you think is causing this problem? (circle one)

Insect A Insect B Insect C

What makes you think so?_____

_____

_____

When you think about it, content validity is not a form of validity at all. It has more to do with the adequacy of the sampling plan than with the inferences you make on the basis of a classroom-based assessment task. Content validity is only one piece of information necessary for evaluating the validity of your inferences. To determine the validity of inferences about student performance, we would have to consider the construct as well as the content sampled.

As you may recall, we said the purpose of classroom-based assessment is to make instructional decisions. Recently, our notions of validity have expanded greatly to include a wider range of variables associated with planning instruction, and with the specific uses to which assessment data are put. New terms are finding their way into the vocabulary of assessment. In the remainder of this section, we will take a brief look at recent conceptions of validity. We believe these alternative views of validity are important because they link the assessment and decision-making processes more closely with instruction. Some of these new views of validity have features that overlap with one another and with the more traditional forms of validity described above. However, they all share a basis in the perspective that valid decision making must be built into the instructional planning process.

Systemic Validity

The concept of systemic validity was proposed by Fredericksen and Collins[15] as an expanded view of construct validity that takes into account the instructional changes that a test engenders. According to Fredericksen and Collins, a systemically valid test is one that changes the instructional system to foster the development of the cognitive skills the test is designed to measure. When an assessment program is systemically valid, instruction begins to focus on the cognitive skills that are the focus of the program, and students then subsequently begin to show improvement in those skills. In other words, instruction focuses on the processes involved in performing assessment tasks rather than on completing specific instances of the task (teaching to the process rather than teaching to the test). Here is an example that illustrates the difference between teaching to the process rather than teaching to the test.

---

[15] Fredericksen, J. R, & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Ms. Wolf is a seventh grade social studies teacher. She uses the *Social Studies for Today* curriculum series. This series provides the teacher with ready made lecture outlines, homework assignments, and end-of-unit tests. Ms. Wolf likes using in class simulation activities that she makes up because they give students lots of opportunities to use the concepts that she feels are important. However, she is disappointed that her students find the activities so difficult. Students always ask her for the correct answer and seem unable to handle the idea that more than one correct answer might make sense in some situations. Her students seem to be very good at answering the items on the end-of-unit tests, but they seem unable to use the facts and concepts in solving real life problems. Ms. Wolf also finds that if she strays too far from the lecture notes provided by the curriculum program, students tend to do poorly on the end-of-chapter tests.

Ms. Wolf began to suspect that the test did not representatively sample the domain in which she provided instruction. When she looked closer, she realized the tests were mostly reiteration of facts. She knew she could get her students to score high on these test just by going over the specific facts included on the test. Ms. Wolf decided to make up her own tests that required students to evaluate, predict, and apply key concepts and principles. She found that students didn't know where to begin. When asked to make an evaluation, they might not provide a rationale for the decision they made. In making a prediction, students tended to summarize the information in the prompt but didn't move beyond this information to tell what would happen next. Or, when they did make predictions, often these were based on conjecture rather than on use of key concepts and principles.

Ms. Wolf began teaching her students some strategies for making evaluations, predictions, and applications. She showed the students that to make an evaluation they need to indicate a clear choice among the options presented and then present a convincing argument for making the choice. She taught the students the difference between this and simply summarizing information given in the prompt.

Similarly, she showed her students that, in making predictions, they should provide a cogent rationale for making a highly probable prediction rather than making a guess based on opinion. Finally, she showed her students how to use the skills in a wide range of social studies contexts. Each new chapter in the text book presented a new context in which to teach her students to use complex intellectual operations. Soon the students began obtaining higher scores on the end-of-chapter tests that Ms. Wolf made.

Assessments that have systemic validity possess two attributes. First, they employ direct measures of the cognitive skills being tested. Second, they employ subjective, or qualitative scoring systems. Direct measures of the intellectual operations we have been talking about in this module would be those that require students to solve problems in social studies or science using evaluation, prediction, and application of key concepts and principles. Indirect measures would be those that simply require students to summarize and reiterate facts, concepts, and principles but never use them in more sophisticated operations.

Subjective scoring requires more training for scorers to obtain consistent results but less inference in determining whether or not students can actually

perform the cognitive tasks. For example, if you rate the extent to which students have used key concepts in evaluation on an essay or interview task, your inferences about student thinking will be more valid than if you simply ask students to use facts on multiple choice and true-false items.

The distinction between direct and indirect tasks may be difficult to grasp at first, but the distinction is an important one. An example from science may provide some clarification:

> In a recent unit on energy, much discussion was devoted to fossil fuels The goal of your instruction was for students to be able to identify the different kinds of fossil fuels and how they are used in daily life. An equal amount of time was spent on the environmental effects of fossil fuel use, concluding with a discussion of some alternative energy sources.

Here is one set of questions you could ask:

1. Name three kinds of fossil fuels:

   1. _____
   2. _____
   3. _____

2. Name three products that are made from fossil fuels:

   1. _____
   2. _____
   3. _____

3. What is acid precipitation? _____

   _____

   _____

4. In the list of energy sources below, circle the ones that are not fossil fuels.

   a. wood b. ethanol

   c. natural gas d. coal

Here is another item you could use:

The use of fossil fuels, especially oil, has increased drastically this century. At the same time the amount of acid precipitation has also increased. What would happen if there was a sudden world-wide decrease in the use of oil and products made from oil? Which of these three outcomes in most likely to happen?

A. There would be a sudden decrease in the amount of acid precipitation in the Northeastern United States and maritime Canada.

B. Other kinds of fossil fuel would continue to create acid precipitation. More coal would be used in factories, so acid precipitation would increase.

C. No change in the amount of acid rain would occur, because acid precipitation would continue to be made from natural sources and through the use of other kinds of fossil fuels.

Write an essay that tells why you made the choice you did. If you think acid precipitation will decrease if we decrease the amount of oil we use, tell why. If you think acid precipitation will increase, tell why. If you think the amount of acid rain will stay the same, tell why.

Your essay will be scored according to the accuracy of the information you use, and the quality of your argument.

The first four items on the previous page are indirect tasks because students are never asked to demonstrate an understanding of the ways we use different sources of energy or of the difference between renewable and non-renewable resources, the stated goals of instruction. The last item is a direct measure of students' use of the operation prediction. Given a set of circumstances (heavy use of fossil fuels, creation of acid precipitation), students must tell what will happen when one variable (amount of oil used) is changed. To respond to the item, students need to use factual information (fossil fuel use contributes to acid precipitation) and principles (the more fossil fuels are burned, the more acid precipitation is created). Any of the three choices could be correct; each could be supported with information that was provided during instruction. Students' use of prediction can be judged on a continuum—or scale—of competence, based on the accuracy of the information and the strength of the argument used.

Scaling implies that a behavior can exist at different values rather than in a dichotomous "exists"/"does not exist" format. All students would be assumed to be able to use information in complex intellectual operations, just that some might do it better than others. This is different from the notion that a student either can or can't

answer an item. Instruction is systemically valid if it moves students along a continuum of competence, regardless of their current level performance.

Educational Validity

With experimental design, issues of validity traditionally have been separated into two categories: *external* and *internal* validity.[16] *External validity* is concerned with generalizability—it asks, "Do the results obtained in one particular situation apply in other situations?" For example, if Ms. Jones notices that her science students improved their organization in writing essays after she began requiring them to outline their class notes, will Mr. Smith see the same result if he requires his history students to do the same?

Internal validity is the fundamental basis for interpreting the results of an intervention—it asks, "Did the intervention make a difference in this specific instance?" For internal validity to be supported, an effect must be demonstrated to be the direct result of a particular procedure. Extending the above example, we can ask the question, "Did outlining their class notes lead Ms. Jones' students to write essays that were more organized?" Or was this observed improvement due to some other cause, such as practice in essay writing in their English classes, the switch to a new science topic about which the students had extensive prior knowledge, or the fact that Ms. Jones simply gave them twice as much time as she previously had to complete the essays?

As the above examples show, these issues of validity need not be limited to tests and research methodology but can be raised in reference to instructional interventions. For example, *social validity* refers to the social value and acceptability of an educational intervention.[17] Socially valid educational programs focus on skill or knowledge that is important to the learner in current or future environments.

---

[16] Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research* . Chicago: Rand McNally.

[17] Kazdin, A. (1977). Assessing the clinical or applied importance of behavior change through social validity. *Behavior Modification, 1*(14), 427-451.
Wolf, M. (1978). Social validity: the case for subjective measurement or how applied behavior analysis is finding its heart, *Journal of Applied Behavior Analysis, 11*(2), 203-214.

Educational validity is a broader term that encompasses all of these different concepts of validity.[18] For an intervention to have educational validity, three criteria must be met. First, a behavior change must occur as a function of the intervention. Second, the intervention must be implemented as planned. And third, the intervention must be beneficial to the student, i.e., socially valid. Each of these criteria is necessary and no one is sufficient by itself. With educational validity we ask, "Did the instructional intervention teach the skills or knowledge it was intended to teach and is that content important to the student in current or future environments?"

Educational validity pertains to the specific inference that, as a direct result of instruction, a student has learned something useful. This is not a trivial or esoteric research issue. For all practical purposes, students are required to attend school. Furthermore, we are currently in the midst of an information revolution. The quantity of skills and information students need to master to function effectively in post-school environments continues to accelerate. But the time available for students to learn has not changed appreciably for over fifty years. Given the time crunch under which teachers and students now operate, how can we justify having students who sit in our classrooms waste their time in activities that fail to teach useful skills and knowledge?

Although educators assume that students learn important skills and knowledge as a function of instruction, the educational validity of our instructional interventions receives little systematic scrutiny, particularly in content areas such as social studies and science. Consequently, teachers are held accountable to a far less rigorous standard than test developers or scientists. Yet, daily classroom instructional interventions can have a far greater potential impact on students than any single test or experiment. The assessment technology typically used in schools provides little information about the actual effects of instruction. Instruction is only loosely aligned with goals, and these, in turn, are rarely analyzed for their functional value to the learner.

The procedures described in this module and in *Training Module 3* are designed to ensure educational validity. The process of *selecting* content is a planful activity linked directly to *interactive instruction.* Together, the content and the use of content comprise the *instructional domain.* This instructional domain is essentially the same as the *assessment domain,* from which *assessment tasks* are sampled. In the end,

---

[18] Voeltz, L. M., & Evans, I. M. (1983). Educational validity: Procedures to evaluate outcomes in programs for severely handicapped learners. *Journal of the Association for the Severely Handicapped, 8*(1), 3-15.

student performance is judged according to how much *relevant content* has been learned, as well as how effectively this content is *used.*

Reliability

Technical adequacy is the term that generally refers to two aspects of an assessment procedure: validity and reliability. We have already talked about validity. Validity has to do with the quality of inferences you can make about students on the basis of assessment results. Reliability is the other side of the coin. Reliability is concerned with the amount of "static" or "random error" an assessment system contains that may decrease the validity of the inferences you make.

Reliability is considered the minimally essential ingredient for determining the worth of a testing or assessment procedure. Usually, reliability is addressed in the context of a specific test, but reliability also is central to the broader process of classroom-based assessment. Therefore, it is critical for you to have a clear understanding of the term.

Often, reliability is interpreted as relating to the stability or consistency of a test or assessment procedure. We might say a procedure is reliable if it yields the same kind of results time after time. However, it is probably more appropriate to think in terms of the construct that underlies the notion of reliability, random error.

What is Random Error?

Whenever we set out to measure something, we have to contend with variables that may affect our measurements but that have nothing to do with the thing we want to measure. Consider this example:

> Jake decided it was time to paint his house. He went to the paint store and ordered 8 gallons of his favorite shade of Sea-Mist green. The paint salesman informed Jake that color had been discontinued and would have to be custom mixed. Little did the salesman know that his paint mixing machine had a defective valve. Jake's Sea Mist green ended up looking more like Prairie Sage.
>
> When Jake got over his disappointment with the color, he settled down to the serious business of painting. He soon discovered that the paint covered much less area than he had previously estimated. Little did Jake know that he had misread the measuring tape when he was estimating the size of the surface to be painted. When he ran out of paint, Jake went back to the store to order more. This time, the new stock clerk mixed the paint. He carefully followed the formula for Sea Mist green, pouring in the exact amounts of mountain green, frost white, and true blue pigments.
>
> Little did the stock clerk know that the measuring cup he used was different than the one used by the paint salesman who mixed Jake's first batch of paint. As luck would have it though, the amount that the stock clerk's cup differed from the paint salesman's cup exactly compensated for the defective valve in the paint mixing machine. So Jake's new batch of Sea Mist green still looked more like Prairie Sage but at least it matched the other 8 gallons Jake had already painted onto his house.

Measurement variables that affected the quality of Jake's painting job included the faulty valve on the mixing machine, Jake's misreading of the tape measure, and the stock clerk using a different measuring cup. None of these factors was known to Jake, and none had anything to do with his ability to apply paint to the house. These were random events that Jake could not have predicted. In fact, if he could have predicted them, he probably would have taken steps to avoid or remedy them. This example illustrates a fundamental rule in assessment:

> ***All measurement systems contain random error.***

Random error is the term used to describe random events that have nothing to do with the quality being measured. **This kind of error does not refer to the mistakes made by a test taker.** Rather we are talking about error inherent in the assessment system. In a sense, we are talking about errors made by the assessment designer, because, to a large extent, random error occurs when an assessment system contains design flaws. Here are some possible sources of random error (i.e., design flaws) in classroom assessment:

Variation in Test Content and Construction

- If an assessment task is too easy, students may not take the task seriously. On the other hand, if it is too hard, students may guess or give up trying. In either case, the task is more a measure of student motivation than learning.

- If an assessment task is not representative of the instruction delivered, students may not have been provided sufficient information to respond. Representativeness of the domain can involve both knowledge forms (specific facts, concepts, and principles in the content) or intellectual operations. The task may be a measure of students' previous background knowledge rather than an indication of the effectiveness of instruction. This problem is insidious because, often the students with the most background knowledge are the ones who tend to achieve highest on classroom measures anyway. The results of the test could suggest that students learned what you intended them to learn with different rates of success (high-achieving students do better than low-achieving students on the task).

- Poorly worded items or instructions may cause students to respond inappropriately. Different students may respond to the task in different ways. In this case, the task may be a measure of students' ability to decipher the meaning and intent of the item rather than a measure of what they have learned.

- If a task is not a large enough sample of the domain, students may not have sufficient opportunity to demonstrate what they have learned. This could happen if a test contains too few items or if a production response prompt (such as an essay or an interview) fails to elicit enough talking or writing. Obviously, one or two word responses provide much less information than extended essay or dialog response.

Variation in Administration

- If non-standard administration procedures are used, students in different groups (for example morning and afternoon sections of a social studies class) might be responding to very different tasks. Similarly, if different administration procedures are used on otherwise comparable tasks presented at different points in time (for example a weekly writing test), students may respond very differently on each occasion. To appreciate the impact of administration procedures, you need only consider the affect of factors such as the amount of time students have to complete a task, access to an open textbook, erasing key words or phrases from the board, or prompting students to "remember what we talked about in class yesterday."

- In any assessment situation, students should be as free as possible from environmental distractions. These could include a wide range of variables, such as excess noise, poor lighting, extremes of temperature, uncomfortable seating, and crowded conditions. Less obvious environmental distractions could be introduced in the assessment situation itself. For example, an assessment task might be administered to part of a class while the teacher presents instruction to the rest of the students. Other distractions may occur from outside the room (for example the custodian mowing the lawn, or messages over the intercom).

- The difference between one-to-one and group administered assessments should not be underestimated. In a one-to-one situation, the student may have more opportunities to ask for clarification, and the administrator may be more likely to provide prompts or extended instruction. In a group situation, students may feel under pressure to finish at the same time as others, or they may be less likely to ask for clarification of confusing directions.

- The skill of the administrator can introduce considerable error in the assessment process. If directions are not read clearly, or if the administrator is unfamiliar with the format of the assessment task, students may respond in unexpected ways.

Response Variations by Students

- If students are not sufficiently motivated to respond to an assessment, the results may not be a true indicator of their learning. Student motivation should be built into the assessment process.

- If there are real developmental differences among students responding to the assessment and these differences are directly related to the thing being measured, then the assessment results may be an indication of these developmental differences rather than an indication of learning. This may be a greater problem in lower grades where greater inter-student developmental differences may be seen, but it could be an issue in later school years if there are real developmental differences in various aspects of conceptual or intellectual development.

- If students are particularly anxious about the outcome of an assessment, they may respond erratically. This is the flip side of the issue of motivation; students may be too highly motivated. Test anxiety may be an issue for specific students whose previous experiences with tests were punishing. Anxiety also may be an issue if students perceive the assessment as a "high-stakes" situation (for example, a large proportion of quarterly grades are based on one test).

- If students lack specific test-taking skills associated with a particular assessment task, they may not respond in an appropriate manner. This issue involves but also goes beyond the notion of being "test savvy." Clearly, there are some test-taking strategies that can affect students performance (knowing when to guess, knowing how to interpret item stems, etc.). However, if students are presented with unorthodox assessment tasks, they may need explicit instructions on how to respond. For example in an essay task that requires students to make a choice between two competing options and then present a defense for the choice, students may need to be instructed that there is no single correct answer. Either choice might be correct, depending on their justification.

- Student physiological variables always can contribute to the random error in an assessment situation. If a student is hungry, tired, or ill, the results of the assessment task would not represent learning as much as how the child feels on the day the task is administered.

Variation in Scoring

- One of the most likely places for error to occur is in the scoring process. Two scorers might use very different standards for judging student responses. Similarly, at any two points in time, a single scorer might use different standards. In each of these situations, the results could indicate differences in the scoring rather than differences in student performance.

- A closely related issue has to do with the standards used within a single scoring of an assessment task. If the standards change part of the way through the process, the results will reflect this source of error. Consider an essay measure that is being rated according to a five-point scale. All

essays would need to be compared with the same set of anchors. If the rater decides that the scales are too complicated and changes them mid-way through the process, then all essays would need to be re-scored with the new scales.

What Does "Reliable" Mean?

All measurement systems contain some amount of random error. When we know that random error is being introduced into an assessment situation, we can take steps to eliminate its source. Unfortunately, we can never be sure we have identified *all* of the sources of random error. Therefore, we can never really be sure whether the results of our assessment represent real observations or the effects of error. Here is an example that illustrates this point:

> Ms. Sparks teaches a morning and an afternoon section of sixth grade social studies. At the end of two weeks of instruction about economic development in Central America, she administered an essay task in both sections of the class. The essay prompt presented three types of economic reforms that a developing country could adopt and asked students to pick one and write an essay that justified the choice.
>
> When Ms. Sparks looked at the essays her students wrote she was pleased to find that her morning class had used a number of key terms she had presented in instruction, but she was disappointed to find that only a few of the students in the afternoon class had used these terms.
>
> Upon further examination of the essays, Ms. Sparks noticed that a number of students in the morning class had simply summarized the attributes of the economic reform they selected but had not presented a rationale for their choice. Most of the students in the afternoon class presented a cogent rationale for the type of economic reform they chose. Finally, Ms. Sparks noticed that one of the students in the afternoon class who almost always writes long, colorful essays wrote only a few sentences.
>
> What should Ms. Sparks make of her results?

Ms. Sparks wanted to find out if her students could use the information she had presented during two weeks of instruction. As it turns out, she'll never know whether the results she obtained on the essay task were influenced by her instruction or by irrelevant variables that had nothing to do with how well she taught the material. If the essay task had been free from random error and her instruction in the two groups was comparable, she should have seen fairly consistent results in both her morning and afternoon groups. The reliability of the assessment task must be suspected.

The term "reliable" means "free from random error." The more a measure is influenced by factors that are irrelevant to the purpose of the task, the less reliable it is. Because all measures contain some amount of random error, no assessment task is completely reliable. Reliability is like happiness: it's always nice to have some, and the more you have, the better off you are.

Decide what kind of error might be present in the following test situations and discuss ways to control it.

1. A weekly spelling test is administered by a substitute teacher.
2. A reading test consists of items that require students to read a passage and answer multiple choice questions.
3. Written expression essay tests are graded according to teacher judgment of communicative competence.
4. The morning division of math class is tested on Monday and the afternoon division is tested on Wednesday; both groups get the same test.
5. Schools in two separate districts have different policies regarding the amount of preparation students receive prior to administration of the State Basic Skills test.
6. A school readiness screening test is administered on the first day of kindergarten.
7. End-of-the-chapter tests for *Your World of Geography* are provided by the textbook publisher. Mr. Jones uses these tests to assess student progress.
8. A state-wide achievement test administered in all eighth grade classes in the state, at the same time each year. Schools often are compared on the basis of the results of this test.

There are two ways to handle reliability in designing and using classroom assessment. First, you can take as many steps as possible to reduce the effects of random error listed above when you construct and administer an assessment task. This is a proactive approach to ensuring reliability.

The second way you can handle reliability is by using the tools of researchers and large-scale test developers to verify that your measures are free from random error. This approach is applied after students have completed the assessment task, when you are scoring their responses. It involves traditional notions of test-retest, alternate-forms, internal consistency and inter-rater reliability. We will discuss these kinds of reliability in the next section.

Strategies for Estimating Reliability

Traditional notions of reliability often refer to various types of reliability. In fact, there is only one type of reliability, either a measure is reliable (i.e., relatively free from random error) or it is not. However, there are a number of types of strategies for estimating the reliability of an assessment task. Some of these strategies pertain more to selection response tests than the kind of extended production responses we have been talking about in this training module, but they will be presented here to provide a full picture of reliability estimation.

There are some potential problems with using these strategies in a classroom-based assessment program. First, the important point to remember is that verifying reliability can only happen *after* students have completed the assessment task. If it turns out that the measure is *not* reliable, the value of the results you obtain is severely limited. Depending on how much random error you think the measure contains, you may not be able to use the results in good conscience. The second potential problem is that each of these strategies requires extra effort above and beyond the normal planning that teachers generally do in classroom settings. For example in some cases, an extra version of an assessment task may be required, or a second person may need to score student responses.

The prudent approach is to take as many steps as possible to ensure reliability during the planning, administration, and scoring processes. On the other hand, it is be better to know your measures are unreliable than to make decisions on the basis of unreliable measures that you didn't know were unreliable. Therefore, the best approach would be to build in strategies for ensuring *and* verifying reliability in an assessment program.

After you have done all you can to control for random error during the design and administration of an assessment task, there are a number of methods you can use to estimate how reliable your assessment actually was. Three of these methods employ numerical procedures with students' responses to the items on an assessment task: Parallel forms reliability, test-retest, and internal consistency reliability. A fourth kind of reliability estimates the consistency and stability of the procedures used to score students' responses, either across time or across scorers. This kind of reliability estimate is called intra- or inter-rater reliability. Each of these methods for estimating reliability is summarized on the following pages.

Estimatng Reliability with Parallel Forms (a.k.a. Alternate Forms or Equivalent Forms)

Description:
- Two forms of the same task are administered to one group of students at about the same point in time.

When Used:
- Two or more forms of the same task exist.

Assumptions:
- Both forms of the task measure the same content or skill.
- Both forms are administered to the same group.
- The tasks are equivalent but not identical.
- The two forms must be of a comparable level of difficulty.
- Administration and scoring is the same for each form.

Caveats:
- Students' performance may be subject to practice effects. That is, their performance on the second version should not be sensitive to the practice they receive by taking the first version.
- A parallel form strategy may be difficult to establish if the domain sampled is too large.

Estimating Reliability with a Test-Retest Strategy

Description:
- Measures the stability of scores between two administrations of the same test. The same task is administered to a single group of students at two different points in time (about a week apart).

When Used:
- When task is a broad measure of achievement not linked to a particular curriculum.

Assumptions:
- Student performance will not fluctuate between task administrations.
- Students do not receive instruction specific to the task between administrations.
- Administration and scoring is the same each time.

Caveats:
- This strategy is sensitive to practice effects or instruction.
- When items are sampled from a small domain, there may be little variation among students and thus lower correlation coefficients.

Estimating Reliability with a Parallel Form and Test-Retest Strategy

Description:
- Two different forms of a task are administered at two different times. At Time 1, half the student group receives form A and the other half receives form B. At time 2, the students receive the alternate form.

When Used:
- To establish reliability of a published norm-referenced achievement test as a pre-post measure over a long period of time.

Assumptions:
- All that apply to Parallel Form or Test-Retest reliability apply.

Caveats:
- Because this procedure is so stringent, a task may be unnecessarily rejected.

Estimating Reliability with an Internal Consistency (Split-half) Strategy

Description:
- Measures the extent to which items contained in a single *test* are interrelated. Based upon the average intercorrelation among items within a test.

When Used:
- When making inferences about students' performance in a broad domain.

Assumptions:
- The test can be divided into two equivalent halves.
- The test contains enough items to split it into two halves.

Caveats:
- Cannot be used on timed tasks where students don't finish all items.
- Provides no estimate of reliability over time or forms.
- Is most useful for tests, not extended production tasks.

Estimating Reliability with an Inter- and Intra-Judge Agreement Strategy

Description:
- Indicates the extent of agreement among or within judges, raters, or scorers over time. It is sensitive only to scoring procedures and ignores other sources of error.

When Used:
- When scoring is subjective.
- When an individual judge repeats a scoring procedure at a later point in time.

Assumptions:
- Scoring is conducted under the same conditions on both occasions. Judges have been adequately trained and scoring conditions are optimal.

Caveats:
- Inter-judge agreement indicates only that a particular outcome is seen by all judges trained to see it. Others not trained in the scoring procedure may or may not observe the same outcome.

Note:
- ***Inter- /Intra-rater reliability is the most valuable and appropriate strategy for estimating reliability in extended production responses. This strategy should become a routine part of an assessment program that uses production responses such as essays, interviews, and projects.***

The exercise below gives you an opportunity to apply these different types of reliability to various testing situations.

Decide which type or types of reliability would be appropriate for each of the following tests:

1. An essay test scored by panel of three teachers on a 5-point scale of creativity and communicative competence.
2. A nationally normed achievement test intended to sample skills in reading, math and language constructed to reflect the content of a variety of curriculum programs.
3. A math computation test to be used in a pre-post format to measure the effects of instruction delivered over a one-month period.
4. Oral reading fluency tests that employ passages sampled randomly from level 5 of the basal reading series.
5. A test of written expression scored by counting the total number of words written.
6. An achievement test used to place students into or remove students from a specialized program on the basis of rate of growth over time.
7. A teacher-made end-of-the-unit test in tenth grade World History.
8. A background knowledge test used to decide who will be allowed to enroll in the International Studies Program.

Creating Prompts for Extended Production Responses

Throughout this training module, we have been talking about the kind of planning you need to do to ensure reliable and valid assessment systems. In this final section we will present some guidelines for translating this planning into the actual prompts to which students will respond. We will talk mostly about written (i.e., essay) response but all of the issues we discuss apply to assessments that require other responses, such as interviews, and graphic displays (for example, maps or diagrams).

Remember, we are interested in looking at student performance in material taught in content classes. Therefore, we will always proceed from an analysis of the knowledge forms (facts, concepts, and principles) and intellectual operations (reiterate, summarize, illustrate, predict, evaluate, and apply) that comprise the assessment domain.

Development of assessment prompts needs to proceed by first considering the intellectual operation and then identifying the knowledge forms which will be used in the response. Following are a number of issues to consider as you develop prompts.

Architecture of the Prompt

Like blueprints for building construction, assessment prompts should have a structure that identifies specific informational content and format. At the very least, prompts should have three parts: (a) introduction, (b) knowledge forms, and (c) intellectual operation stem. The opening sentences should establish the context and setting of the problem, the broad area of content (topic), and the context of the task. At least one subsequent sentence should introduce a range of knowledge forms to be used within this specific context and content. Finally, the response demand made upon students should ensure they are performing the correct intellectual operation in their answers.

With some intellectual operations, the architecture may be more elaborate. For example, with evaluation prompts, the middle paragraph introduces the knowledge forms and may present (at least) two sides of an issue: advantages and

disadvantages, similarities and dissimilarities, positive and negative influences, etc. The purpose of the evaluation stem is to prompt the student to consider both sides and make a choice between these alternatives. Or, with a prediction item, the middle paragraph may need to include a time sequence to convey the correct trajectory of events so that the eventual prediction stem can depart from a well-established point.

Pivotal Words in the Stem

Specific and pivotal words are the most important influences in the prompt. They should be chosen with care because they are intended to lead the student to a specific intellectual operation. Generally, these words are verbs, though they may be adjectives (with illustration) and verb-objects. Following are examples of pivotal words for each type of intellectual operation, with the key phrases underlined:

- **Reiteration:** *Repeat the exact words, recite, state the definition, give a verbatim* description...

  Example: "Recite the preamble to the Bill of Rights."

- **Summarization:** *Paraphrase* the content, *retell* what you read, *summarize* what was said, *describe* the main issue...

  Example: "Describe the main problem engineers encountered when they first tried to extract oil from shale."

- **Illustration:** Provide an *example*, present a *comparable issue*, relate an *analogous incident*...

  Example: "Provide an example from your own experience of the consequences of procrastinating."

- **Evaluation:** *Decide* which *alternative*, determine the *correct choice*, *consider* which *option*, *compare* and *determine*, *select* one and *justify*...

  Example: "Of the alternatives to the use of fossil fuels listed above, select the one that you think would work the best and justify your answer."

- **Prediction:** Tell *what will happen, make a prediction, guess an outcome,* describe *subsequent events*...

  Example: "If all the plankton die in a lake, describe what will happen to the other organisms living in the lake."

- **Application:** *Explain the outcomes, give some reasons...*

  Example: "Use what you have learned about plate tectonics to explain the existence of the Cascade mountain range."

Equality of Choices

A potential source of error, and thus a threat to reliability and validity, is the amount of "bias" contained in a prompt. Students should not be tipped off by the wording that one type of response is better than another. They should be able to demonstrate what they have learned by making effective choices and arguments on their own. The prompt must be worded so that all choices which are (implicitly) embedded within the text are equal. The prompt should not convey the message that one choice is better than another. Although this requirement sounds obvious, a number of subtle influences may exist that make one choice more relevant or "answerable" than any other. Following are some examples:

- In describing the content (knowledge forms) and context, more information may be provided for one specific concept than another. For example, if an essay about two forms of economy (command and market) had a disproportionate amount (and specificity) of information about market economics, students' answers may be influenced.

- The breadth of the concepts may be different, so that one choice has fewer opportunities for an extended answer. An essay may cover two biomes (rain forest and tundra) and the natural resources that are converted to human usage within each. Clearly, fewer natural resources are exploited from the tundra, thereby prompting students to answer more often with references to rain forests.

Instructional Relevance

Like all end-of-unit assessments, it is often (and incorrectly) assumed that the information covered on the test has been addressed in the course. Very little research, however, has been completed on this topic from an instructional viewpoint. Rather, the research that has been done has focused on the overlap of the curriculum and published achievement tests.

The issue, then, is how closely the text book is aligned with instruction delivered in the classroom. And, with the framework we have provided (using a

two-dimensional analysis of knowledge forms by intellectual operations), this issue becomes much more complex than just counting the number of times a vocabulary word or informational item is included in instruction and on the test. Rather, it is the specific knowledge form and its use as an intellectual operation that must be clear and consistent across these two dimensions of instruction and assessment. For example, a concept such as fossil fuels may be emphasized heavily within instruction using "prediction" (of positive and negative consequences from heavy dependence upon fossil fuels); however, if the end-of-unit test contains items requiring students to give examples (illustrations) and then summarize their uses, the overlap between instruction and assessment is negligible.

Administration Format

A key issue to be addressed when assessments are administered involves the *level of prompting* and the *task demands.*

Level of prompting, or the amount and kind of supplemental help that students receive to organize their response, can be great or absolutely minimal. For example, the extended text for a prompt may be read to students by the teacher and key words or phrases highlighted along with verbal prompts to recall the content of instruction (e.g., "you need to think about the part of the chapter that covered the nitrogen cycle...remember, how a certain chemical catalyst got the cycle started..."); this kind of administration would be heavily prompted. In contrast, the teacher may just distribute the prompt with the simple directive for them to read the directions and begin when they are ready to write; these procedures would reflect almost no prompting.

Task demands refer to the mode of response required of students. Because written essays leave a permanent product that can be scored and evaluated later, they are easy to use. However, they also require students to somewhat facile in writing. If the purpose of the measure is student performance with a specific intellectual operation, then their response should not be unduly influenced by the manner in which they respond. For example, a student may have exhibited much more (or less) knowledge simply because of the task demands (e.g., writing out their response instead of talking it out). The outcome, then, must be considered as a mix of the intellectual operation and the task demands; it is impossible to sort them out separately.

An issue related to task demand and level of prompting has to do with the amount of time and structure students are provided to prepare their responses. For example, consider the different kinds of responses you could expect in the following situations:

- Mr. Jock gives his social studies students a "pop" quiz about the Civil War the first thing Monday morning.
- Ms. Shine tells her students, "Tomorrow we will have a short essay test about the important events that led to the Civil War."
- Mr. Mixx allocates 30 minutes for students to make an outline prior to writing an essay about plate tectonics. He encourages students to review their notes on sea-floor spreading in making this outline.
- Ms. Moon hands out an essay prompt and instructs students to begin work. This event occurs during the last 15 minutes of a class in which the students watched a movie.
- Mr. Miller presents a list of questions to his students in his health class at the beginning of the period. The student in the first seat of the of the row by the door answers the first question, the student behind her answers the next question, and so on until all students have answered a question.

Scale of the Response

Student responses may be scaled either quantitatively or qualitatively; the degree to which students know how they will be judged may influence how they perform. Actually, arguments can be made both for telling them in advance and for withholding such information from them; the choice depends upon the purposes of the assessment.

For maximal instructional application, students should be told in advance or as part of the prompt, how their performance will be evaluated. For example, they may be directed to "use as many of the important concepts" or "link the important concepts into a major principle" because that is what is valued. Students may even be given the actual qualitative scale to use in constructing their responses (so they can see the anchors for a low response versus a high response).

For maximal generalizability to a larger context or group (e.g., across time, teachers, classes, or grades), students should not be told anything about the scoring

systems but to simply do their best. In these instances, the assessment task should reflect the fact that in generalized settings, students will encounter a variety of prompts and response demands. They would need to be able to extract the key elements of the task and respond appropriately.

Explicit Reference to Elaboration

Generally, students receive relatively little practice in responding to writing or speaking tasks. Therefore, it is very likely that an extended essay using a problem-solving situation will result in very narrow distribution of scores (be centered near the bottom and have only a few high performers). Students probably will need to be explicitly directed to "justify the answer," "give as many reasons as possible," "provide as much supporting detail from the chapter as they can," etc.

These extra prompts should communicate to students that they need to do more than write a minimal response and then stop writing. If extended response items are used regularly in a classroom and students are taught the criteria by which their responses will be judged, this issue will tend to diminish over time.

Response Strategies within Student Performance

In every facet of classroom interactions, students need strategies to perform successfully. For example, when teachers are lecturing, these strategies may involve note-taking. During independent reading, students may need to rely on any of the various SQ3R-type strategies to help them meaningfully interact with the text. A host of test-taking strategies have emerged over the years to help students perform better. It is this last issue which has bearing in the use of extended production and essay tasks, particularly in creating a prompt that generates a response which is capable of being evaluated reliably and validly.

Student responses to essay tasks may be influenced by a number of components such as: knowledge of the content, background knowledge, and writing process skills. To ensure the task can generate reliable estimates of performance, other influences need to be controlled. Following are some suggestions that should help focus the task on the knowledge forms and the intellectual operations of interest.

- Try to include enough broad and commonly known background information so that all students have an equal chance of responding only to the content of the domain.

- End the prompt with directions on how the response should look and even include some qualifying criteria. For example, students should be explicitly told that, although they should write legibly, their answers will **not** be scored for penmanship or spelling; if they don't know how to spell the word, do the best they can. Or, to prevent students from including most of the information from the prompt, students should be directed to "focus on the content of instruction and use reasons from their knowledge of their subject, not just the information from the prompt."

- Encourage students to take a few minutes before they begin writing to review the prompt and plan their response. A planning sheet may also be used to let them organize (and even outline) their ideas before they begin writing.

- Direct students to work the entire time allotted; they should be explicitly told that, even though they may finish early, they should read their response to make sure they have adequately answered the question.

- A running time prompt may be needed to help students pace themselves so they do not spend too much time on any single facet of the response. For example, if an evaluation problem requires students to compare and contrast two events (a summarization task), they may need to be prompted to move on to the part of the task requiring them to make a decision and support it with specific criteria.

- Provide a motivating statement that allows everyone to perform with equal diligence. Although motivation is a key aspect of student performance for all students, it is very difficult to quantify and control it; the most typical strategy is to tie performance on the task to grades in the course, which may be exactly the wrong way to proceed. For example, if students are told that their answer will be used to form half their grade, some students (e.g., those for whom grades are not motivating, those with such bad grades entering into the task that they are flunking regardless of how well they do on this task, those who have performance contingencies on their grades at home) may respond very differently, not as a function of their knowledge, but simply because of the differential effects of the contingencies.

In summary, the prompt should be directed to the specific knowledge forms and intellectual operations of the content and eliminate, as much as possible, influences from background knowledge, administration context that sets the occasion for differential responding, and response process skills (i.e., writing skills).

Embedded Scores within the Prompt

Although it is difficult to ascertain how students will respond in advance of the task, it may be critical to attempt the range of responses that are anticipated by the evaluator. That is, a good way to make sure the prompt is working is to respond to it directly, attempting to write an answer at all levels of quality that might appear within the students' responses. By writing out an "answer key," it may be possible to identify glitches and points of confusion that are part of the prompt. This answer key needs to establish clear differences in a qualitative scale that other (experts) in the field would agree represent differing degrees of performance. And importantly, it needs to establish the extreme scores (the least acceptable and the most elaborate answer).

At some point, the evaluator may need to distinguish between a legitimate score of zero and missing data. For example, if a student responds to the prompt but is incorrect and illogical, a score of zero may be awarded; however, with a response that is on a different topic or clearly represents a misunderstood prompt, it would be more accurate to consider this response as missing (the task was never really attempted). In the end, this trial scoring key may be used either to frame anchors on a qualitative scale or provide prompts to the evaluator on what to identify in student papers when constructing/selecting exemplar papers (range finder papers).

Summary of Prompt Design Strategies

In this section the following nine strategies are presented to help teachers construct adequate extended response prompts:

- Architecture of the Prompt
- Pivotal Words in the Stem
- Independence of Choice
- Instructional Relevance
- Administration Format
- Scale of the Response
- Explicit Reference to Elaboration
- Response Strategies within Student Performance
- Embedded Scores within the Prompt

The major purpose of these extended production response assessments, including student essays, is to allow as much flexibility in constructing an answer so that both student process and product can be evaluated. Process refers to the manner in which students reflect upon and construct meaning using specific knowledge forms and intellectual operations. It refers to diagnostically looking at their "thinking" and is best considered a criterion referenced view. That is, what misconceptions appear to be present? How are different concepts elaborated? What intellectual operations reflect manipulation of the knowledge forms? In contrast, product refers to the outcome, however it is attained. This view is likely to be either norm- or individual-referenced, with the respective purpose being to show relative position or change over time.

Regardless of the focus on process or product, the response needs to interpretable relative to instruction and not unduly influenced by extraneous factors. The list of strategies above should help create worthwhile tasks that are truly classroom-based. If all these suggestions are followed, the inferences made from the data are likely to be more reliable and valid.

Appendix A

Examples of Prompts for Extended Production Responses

Directions for Administering Essays

1. Prepare students for the task by telling them that this is a chance for them to think about what they have learned during the previous one or two weeks of class. Use broad topic labels to direct students' attention to the general content of the essay. For example, you might use wording such as this for a science essay task:

   For the past two weeks we have been talking about fossil fuels. We have discussed the characteristics of fossil fuels and we have talked about the positive and negative points of burning fossil fuels. Today you are going to have an opportunity to use what you have learned about fossil fuels to solve a problem.

   Let students know that there is no single correct answer. The essay prompt will offer two or more choices and any may be acceptable. Tell students that the essays will be scored according to the accuracy of the information students use and the logical soundness of their arguments. Here is some sample wording from a social studies essay prompt:

   You will be asked to choose the best option. Any of the choices with which you will be presented may be acceptable. Your essay will not be scored according to whether or not you pick the "correct" plan. Rather, it will be scored according to how well you present a logical argument to defend your choice, using information you have learned in the past two weeks about forms of government.

2. Read the prompt aloud, with students following along.

3. Ask students if there are any questions. Try to avoid prompting their answers. Give general factual information as needed but don't give rationale for either choice. Remember, this is a form of a test.

4. If needed, remind students to turn the page over after they have made a choice. The essay will be written on the back of the page. Be sure everyone turns over and starts writing.

5. Allow about 15-20 minutes for students to read the prompt and write an essay. If all students are done in a shorter period of time, end the activity.

6. After the essays have been collected, you may wish to use the activity as a stimulus for review or discussion.

First Name	Last Name						Date	
Teacher	Period	1	2	3	4	5	6	7

The Knight and the Serf

This is a story about small kingdom in Central Europe during the Middle Ages. The vassal who rules this small kingdom is Lord Martin. Lord Martin is a mean man. Everyone has to pay him very high taxes. All the serfs in the kingdom work very hard tending Lord Martin's crops but Lord Martin doesn't take good care of them when they need help. He sends his knights off to fight battles but he doesn't give them enough swords. Nobody likes Lord Martin the vassal.

Two people who live in the kingdom have decided they can't take it any more. One of these people is a knight named Sir John. The other person is a serf named Harold. Sir John and Harold have decided they do not want to be loyal to their vassal any longer. They will not pay Lord Martin any more high taxes. Harold the serf will not work hard in the fields and Sir John the knight will not fight any more battles.

Here are two things that could happen if Sir John and Harold stop being loyal to Lord Martin. Place an X beside the statement you think is most accurate.

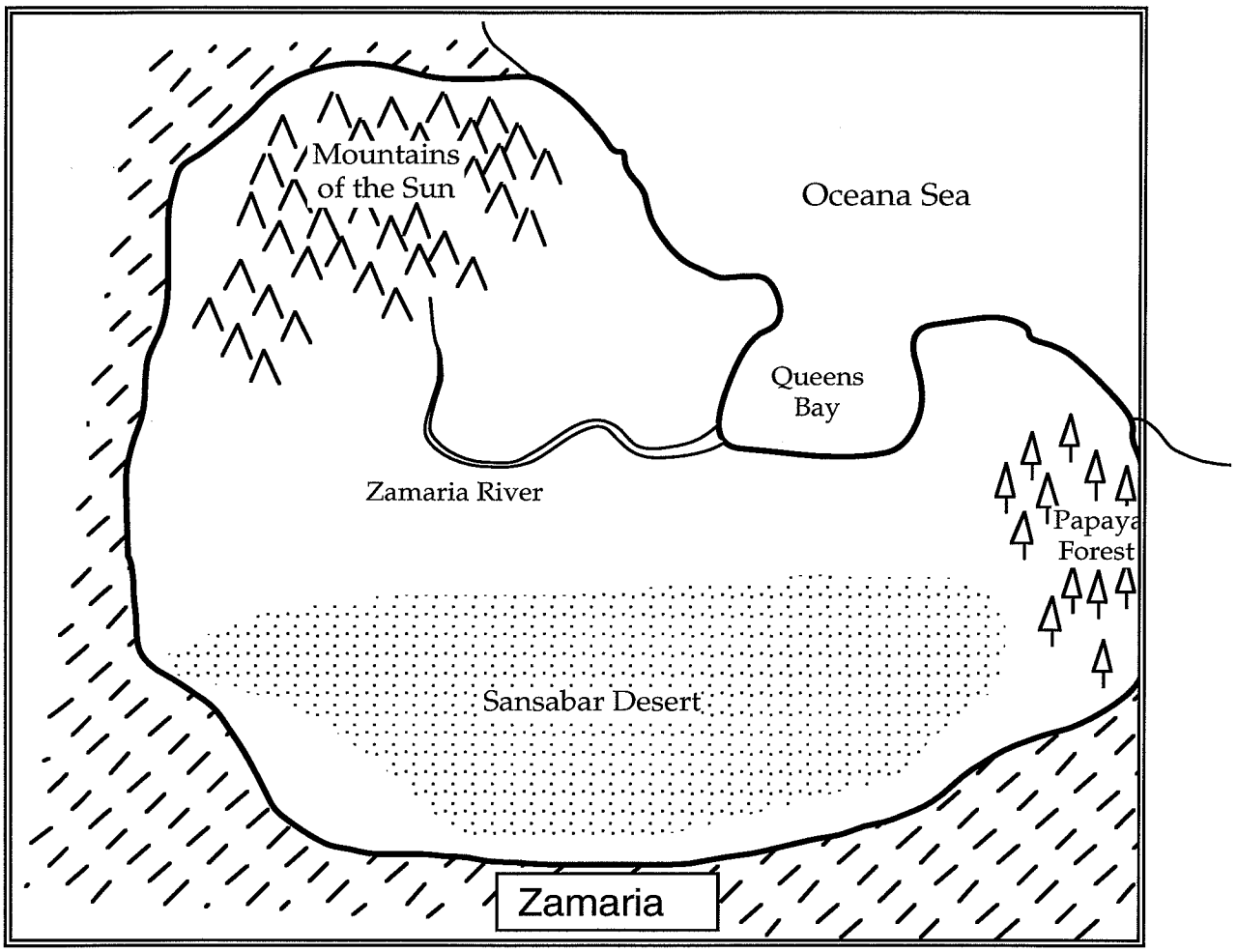
_____ Life will be easier for Sir John the knight than it will be for Harold the serf.

_____ Life will be easier for Harold the serf than it will be for the Sir John knight.

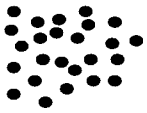
Tell *why* you made your decision. Tell what information you used to make your decision. If you think things will be easier for the serf than for the knight, tell why. If you think things will be easier for the knight than for the serf, tell why.

ZAMARIA

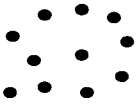
Zamaria is a land that time forgot. No humans have lived there for many centuries. Recently a team of explorers rediscovered Zamaria and now people want to settle there. It is important to know where people will settle so that railroads, sea ports, and cities can be planned. Here is a map of Zamaria. Look at the map and decide where you think people will settle.



Show where you think people will settle by drawing dots to show population density. In areas where you think many people will settle, draw many dots, like this:



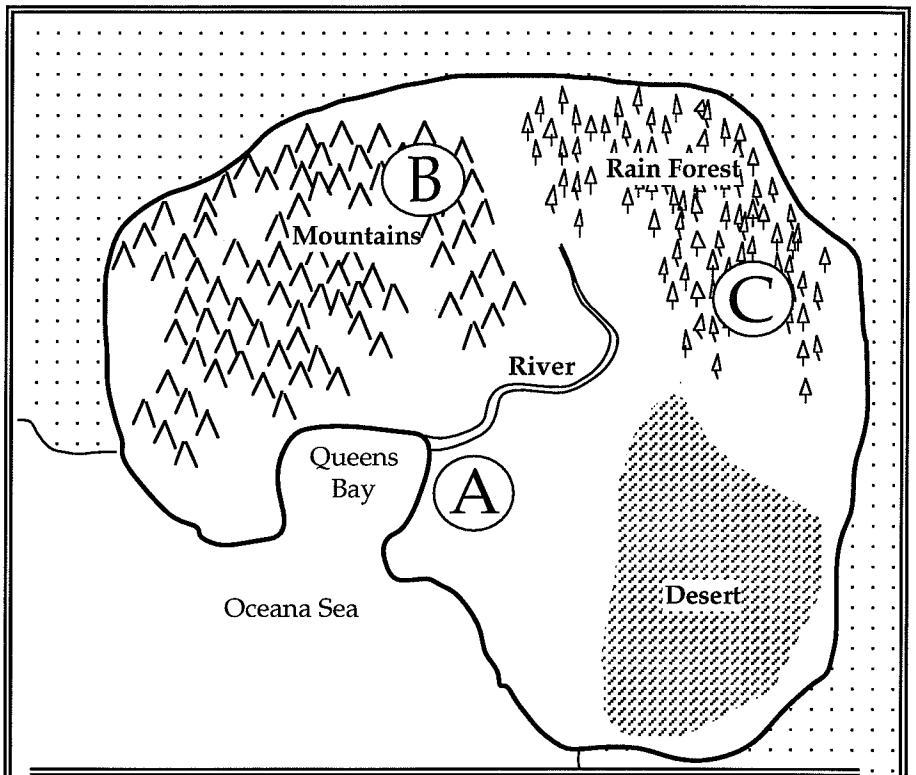
In areas where you think fewer people will settle, draw fewer dots, like this:



In the areas where you think almost no one will settle, don't draw any dots.

Write a short essay that tells why you completed the map the way you did. Tell why you think people will settle in the areas where you drew many dots. Tell what things affect where people settle.

NEW LAND



LEGEND			
	Location		
	A	B	C
Annual Rainfall	4"	30"	70"
Temperature Range	20° to 105°F	-10° to 85°F	50° to 95°F
Elevation	10 feet	4500 feet	500 feet
Natural Resources	minerals	oil	lumber

A team of explorers recently discovered a place that time forgot. Here is a map of this new land. There are high mountains, a rain forest, and a desert. A large river runs through the center of this new place and drains into the ocean. Now that this new land has been discovered, people will want to settle here. Soon towns and cities will develop. Three areas where a city could grow are shown on the map. These areas are labeled with A, B and, C. The legend gives information about the climate, elevation, and natural resources in each area. Where do you think a city will grow? Use what you know about climate, elevation, and natural resources to decide where a city will develop. Mark the blank beside the letter of the area where you think a city will develop.

_____ A _____ B _____ C

Write an essay that tells why you think a city will develop in the area you chose. If you think a city will develop in area "A," tell why you think so. If you think a city will develop in area "B," tell why you think so, or if you think a city will develop in area "C" tell why you think so. Use information from the map and legend to give reasons for your decision.

COSTA HERMOSA

For the last week, you have been learning about the things that affect the economy of a country. Now you will have a chance to use what you have learned to help the people of Costa Hermosa.

Costa Hermosa is a small poor country. There are few roads and no railroads or airports. The best farm land in Costa Hermosa is owned by a few wealthy families. The rest of the people of Costa Hermosa have no land. They grow food to eat in small community gardens but no one has enough land to make a living farming. There are a few factories in Costa Hermosa but they are run very badly. The people who own the factories live in other countries and don't care about making the factories run better. The government of Costa Hermosa has never been stable. The military keeps trying to take over the government. Each state has different rules for electing officials. Many of the people who run the government are corrupt. They take bribes from the rich families who own the land and factories.

Pretend you have just been elected president of Costa Hermosa. You want to make things better for the people of your country. You have to decide what kind of reforms to enact to help the economy. You could enact land reforms to give the poor people enough land to have farms. You could have the government take over the factories to make them run more efficiently. You could make laws that would lead to a more stable government by getting rid of corruption and making the election rules the same in every state. Decide which reform you think is most important. Place an X beside the reform you would enact first.

_____ Take some land away from the rich and give it to the poor.

_____ Have the government take over the industries.

_____ Enact laws to create a stable government.

Write an essay that tells *why* you made your decision. Tell what information you used to make your decision. If you would take some land away from the rich and give it to the poor, tell why. If you would have the government take over the industries, tell why. If you would make laws to make the government more stable, tell why.

WELCOME TO NEWTOPIA

For the past week you have been learning about fossil fuels. Now you will have a chance to use what you have learned to help the people of Newtopia.

Newtopia is a small planet in a galaxy not far from here. It is covered with large trees, the sun shines most of the time and a steady wind blows all year long. People on Newtopia cook and heat their homes with wood from the trees and use animals to do farm work. Gathering firewood and taking care of the animals is hard work so they have very little time to relax and enjoy the beauty of their planet. Recently, scientists have discovered large deposits of fossil fuels on Newtopia. There are oil reservoirs, coal seams, and natural gas deposits.

Some of the people on Newtopia would like use these fossil fuels to heat their homes, and cook their food. They also want to build refineries to make fuels to power tractors, cars, and trains. These people want to make life easier so they will have time for the finer things, like art and music.

Other people on Newtopia do not want to use the fossil fuels. These people are afraid of what will happen to their planet if people start burning coal, oil, and natural gas. This disagreement between the people who want to use the fossil fuels and the people who don't want to use them has caused many arguments and the people of Newtopia don't know what to do.

What do you think the people of Newtopia should do? Should they use the fossil fuels to run their cars and heat their homes or should they leave the fossil fuels in the ground and keep things the way they are now? Place an X beside the statement that tells what you think the people of Newtopia should do.

_____ Use the fossil fuels.

_____ Leave the fossil fuels alone.

Write an essay that tells *why* you made your decision. Tell what information you used to make your decision. If you think the people of Newtopia should use the fossil fuels, tell why. If you think the people of Newtopia should leave the fossil fuels alone, tell why.

CONTINENTS ADRIFT

You are a space geologist in the 25th century. Your job is to describe the geology of far away planets. Recently, a small planet has been discovered in a star system in the Milky Way. You have been asked to help describe it. No one has been to the planet but it has been photographed by satellites. There are five large continents and three large oceans on the planet. You know the continents in one area of this planet are moving but you don't know what kind of movement is happening. Here is the information that has been collected from satellite photos:

1. In the middle of the oceans, there are high ridges.
2. Along the west coast of the biggest continent there are high mountains. Some of these are active volcanoes.
3. Near the coast of one continent there are long narrow regions where the ocean floor looks like a deep valley.
4. All around the edges of the largest ocean there is a ring of volcanoes and earthquake zones.

Decide which kind of continental movement you think is happening on the planet. Decide if you think the movement is happening because of spreading, colliding, or some combination of spreading and colliding. You also have to think about other kinds of movement that might be happening. Maybe it's neither spreading nor colliding. Decide which kind of plate movement is happening. Place an X beside your choice:

_____ spreading

_____ colliding

_____ combination of colliding and spreading

_____ neither kind of movement is happening

Write an essay that tells *why* you made your decision. Tell what information you used. Tell how you decided what kind of movement is happening. If you picked spreading, tell why. If you picked collision, tell why. If you think a combination of spreading and colliding is happening, tell why. If you think neither of these kinds of movement is happening, tell why.

THE SCIENTIFIC METHOD?

Your friend Martha, has lots of plants. She wants to find out how to make her plants grow better. She wants to use the scientific method, but doesn't know what to do next. Here are some things she thinks might affect the way plants grow:

1. The amount of sun light they get.
2. The amount of water they get.
3. The kind of nutrients in the soil.
4. The temperature of the soil.

Write a short essay that tells how you would help Martha find out how what makes plants grow better. Tell how you would conduct an experiment. Tell why the information you would collect is important to help solve this puzzle.
Use the back if you need more room to write.

WHAT IS IT?

Imagine that you are a scientist in the year 2010 and you work in a research laboratory. Your job is to classify living things according to similarities in their body structure and other basic features. Right now you are trying to figure out what to do about an organism that was discovered recently on a remote island in the Pacific Ocean. Here is a description of the organism:

1. It has many cells.
2. It eats other organisms.
3. It can make its own food by using energy from the sun.
4. It can get food by breaking down waste products and dead tissues.
5. During reproduction, it produces spore cases.

Decide what kingdom you think this organism belongs to. Tell why you think it should be classified that way. Tell the features of the organism that make it a member of the kingdom you chose. Also, tell why you think the organism can't be placed in another kingdom.

PLANT OR FUNGUS?

When you go walking in the woods you see many different kinds of living things. You might see squirrels, birds, ferns, trees, mosses, and mushrooms. Sometimes, it is difficult to tell what kingdom an organism belongs to. For example there are many kinds of fungi that resemble plants and some plants look like fungi.

Write an essay that compares the plant and fungi kingdoms. Tell what makes plants different from fungi. Also, tell how fungi are similar to plants.

THE GOVERNMENT OF WALLOO

You live in the small country of Walloo. Walloo is located in the far northern latitudes. It is surrounded by mountains on three sides. A large river runs along one side of the country. Walloo has a large army, but has not fought a war in fifty years. A few people in Walloo are very rich. Most people work long hours to buy food, shelter, and clothing.

The king of Walloo died last week without leaving an heir. A group of army officers announced that they will form an oligarchy if two-thirds of the people don't choose another form of government in a nationwide vote to be held next week. To the east, the dead king's brother rules a country with a command economy. A country to the west shares the same language; it is ruled by a dictator. Walloo trades most of its goods with a poor democratic country to the south.

You are an influential politician in Walloo. Tonight you will be a guest on a popular television program to tell what type of government would be best for Walloo. Write a short essay that tells what form of government you recommend. Tell why you chose this form. Use the back if you need more room to write.

WHO WILL LIVE IN THE NEW COLONY?

The year is 1625 and you live in England. You are planning to move to North America. Some wealthy business people have agreed to give you enough money to pay for your journey and start a new colony near Plymouth, Massachusetts. In exchange for this money, you have agreed to send profitable goods back to England. You have hired a ship with an experienced captain and crew. It can carry 75 other people, a few farm animals, and enough supplies to last 6 months after you land in North America. Now, you must select the people you will take with you to start the new colony. Many others have heard about the colonies in North America and there are more people who want to go than you have room for on the ship. The trip will be difficult, so you must think carefully about who to take with you.

1. To select the best group of people to start a new colony, you must consider many different kinds of information. Here is a list things you might want to know about the people you choose. Decide how important each piece of information is for deciding who to take with you. Put the letter of the most important piece of information in the box next to the number 1. Put the next most important thing in the box next to the number 2. Put the letter of the third most important thing in the next to the number 3, and so on until you have identified the five most important things to consider when deciding who to take with you to North America.

   A. Does the person have experience as a soldier?
   B. What is the climate of Massachusetts like?
   C. How loyal is the person to the Queen of England?
   D. Is the person single or would they bring a family along?
   E. What are the person's religious beliefs?
   F. How many different languages does the person speak?
   G. Is the person good at making things by hand?
   H. Is the person a good hunter?
   I. How wealthy is the person?
   J. Is the person good at reading maps?

   ☐ **1. Most Important**

   ☐ 2. Second Most Important

   ☐ 3. Third most important

   ☐ 4. Fourth most important

   ☐ 5. Fifth most important

2. In one or two paragraphs tell how the information you identified as **Most Important** will help you decide who to take with you to start a new colony in North America.

Appendix B

Example of a Qualitative Scoring System

Scoring System for Evaluation Essays

In *Training Module 3* we presented an example of a scoring system for rating the quality of students' use of the intellectual operation, "evaluation." Here we will present a more elaborate version of that system. As you now know, the prompt that is used to elicit student writing will directly affect the quality of students responses. For example, in making evaluations, students must choose from among at least two options (although more may be presented) and then present a rationale for making the choice. If the choices are clearly presented in the prompt, the task is drastically different than if the options from among which students are expected to choose are only implied in the prompt.

The "Mondo Bondo" prompt that was introduced in *Training Module 3* is shown on the next page. Compare this prompt with the "Newtopia" prompt shown in Appendix A of this module. Both of these prompts deal with the same content and key concepts (related to fossil fuels). The Newtopia prompt requires students to make a selection from among two choices (develop fossil fuels versus leave fossil fuels alone) and then write an essay defending the choice. The Mondo Bondo prompt presents the exact same information, but students are required to tell which choice they made and then write a rationale.

We could think of these two tasks as differing in difficulty. The Mondo Bondo prompt may be more difficult because it requires students to produce more information than the Newtopia prompt. We could also think of these prompts as differing in the extent to which they introduce random error into the process. Remember, the more random error a measure contains, the less reliable it is. We might expect two scorers to reach a higher level of agreement in scoring Newtopia essays than in scoring Mondo Bondo essays because the range of options available for students to write about has been limited to two.

On the other hand, maybe not. Here is an example of a Newtopia essay and a Mondo Bondo essay written by a two sixth grade students. Can you tell which student wrote to the people of Newtopia and which student wrote to the people of Mondo Bondo? (Answer is at the bottom of the next page.)

Student A

People should leave the fossil fuels alone. If the sun shines and the wind only blows a little bit they shouldn't half to heat their houses, except for once in a while. People should use fossil fuels because if they cut down all the trees they won't have any oxygen so they will all die. Also they need transportation and they need oxygen to breathe. they could get more things from fossil fuels to do more things. If the people don't get some rest besides when they sleep they will become lazy and just quit. They also could run out of fossil fuel, then they would half to use the trees for heat and every thing they need and they would die unless they plant more trees every day and soon they won't have any more trees to use. If something happens to their planet they will all die.

Student B

Well there are many good things and bad things about fossil fuels, Well for me I would use the fossil fuels for good things like, heating houses, or putting gas in cars or even putting coal in trains. Hold on you don't know any thing about trains or cars never mind O.K., lets just talk about heating houses or just useful products but if you would like your Island to be kept beautiful & still have beautiful skies maybe you should just make some fossil fuel for like making plastic, cause you see, fossil fuels pollute the air and kill animals & plants but if you use less fossil fuels then maybe damages like killing plants & animals & polluting the air may not be so sever. May they just no kill as much. Well what would ya like, huh, Well bye. Now I need to go to work.

The flow chart scoring system shown on the next three pages is designed to be used with two types of prompts for essays designed to elicit the intellectual operation "evaluation:" those in which the choices are explicit and those in the options are not clearly delineated.[19] This scoring system was based directly from the analytic scoring anchors previously shown in *Training Module 3*. As you can see, though, the flow chart takes into account students' effectiveness in presenting both sides of an argument as well as providing a rationale for the decision made. How would you score the two essays shown above with this system?

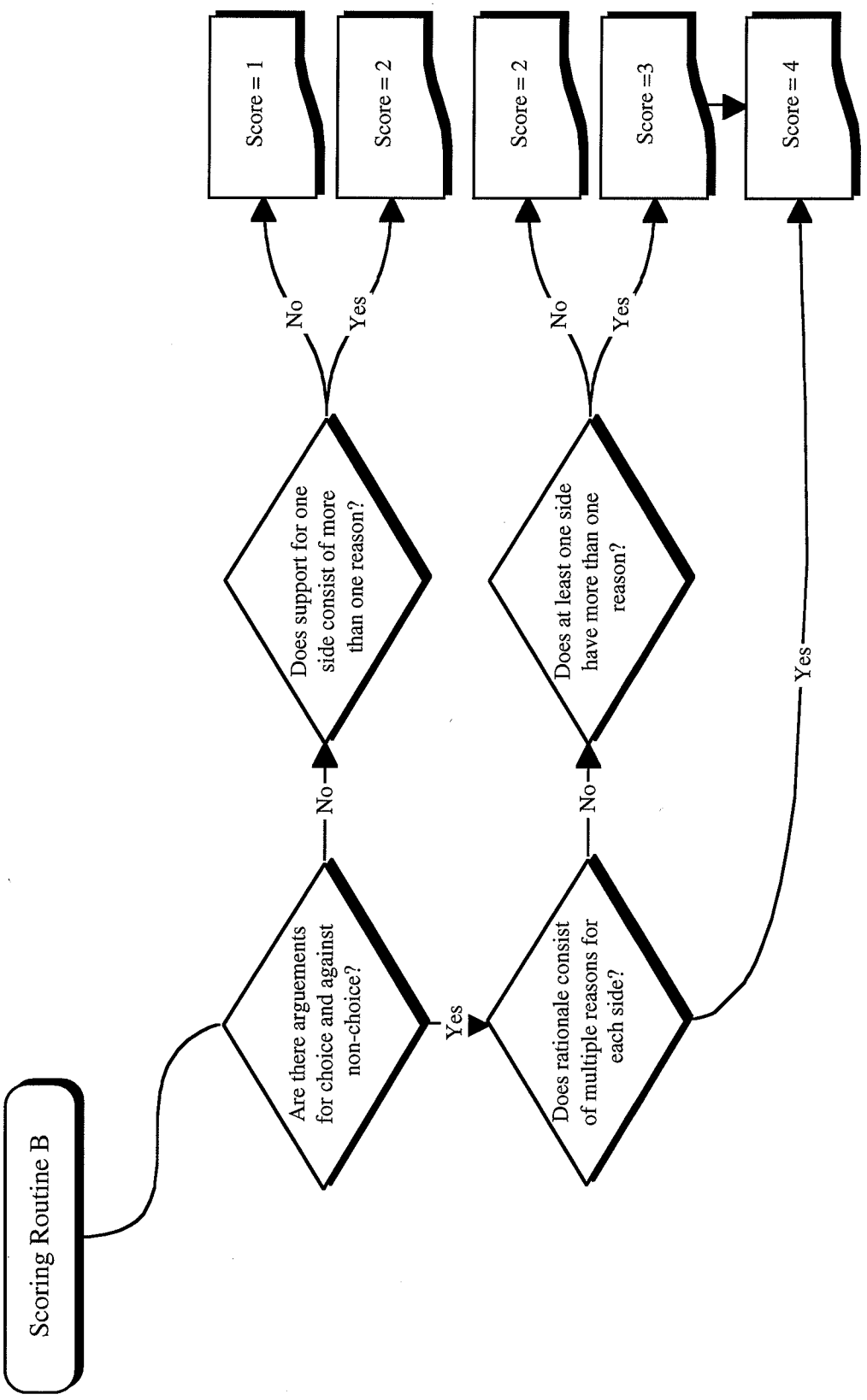
Following the flow charts are examples of various scores obtained with this scoring system to show how it is used. See if you agree with the scores given.

Student A wrote to the people of Newtopia.
Student B wrote to the people of Mondo Bondo.

[19] Nolet, V., Howard, L., & Tindal, G., (1993). *Procedures for scoring evaluation essays.* Unpublished manuscript. University of Oregon, Eugene, OR.

Start
Does the prompt force a choice?
No
Is a decision, or recommendation presented?
No
Score = 0
Yes
Is the choice stated clearly?
No
Is the choice supported with a rationale?
No
Score = 0
Yes
Go to Scoring Routine A
Yes
Go to Scoring Routine B
Yes
Is the essay an attempt to respond to the prompt?
No
Score= 0
Yes
Go to Scoring Routine A

Scoring Routine A
Is the choice supported with a rationale?
No
Score =1
Yes
Are there arguements for choice and against non-choice?
No
Does support for one side consist of more than one reason?
No
Score = 2
Yes
Score = 3
Yes
Does rationale consist of multiple reasons for each side?
No
Does at least one side have more than one reason?
No
Score = 3
Yes
Score =4
Yes
Score = 5


Scoring Routine B
Are there arguements for choice and against non-choice?
No
Yes
Does support for one side consist of more than one reason?
No
Yes
Score = 1
Score = 2
Does rationale consist of multiple reasons for each side?
No
Yes
Does at least one side have more than one reason?
No
Yes
Score = 2
Score =3
Score = 4

The following essays were all written in response to the "What Is It" prompt shown in Appendix A.

Score of 5

This student presented a cogent argument but contains factual errors. However, because the scoring system focuses on students *use* of information, this essay can still receive a score of "5."

> I think this organism is a plant because a plant has many cells, eats lots other organisms, can make its own food by using evergy from the sun., can get food by breaking down waste products and dead tissue, and during reproduction it produces spore cases. The organism couldn't be an animal because an animal can't get food by using energy from the sun, a protist has a one cell structure, monerans do not produce spore cases during reproduction, and fungi doesn't eat other organisms. For these reasons I hope you understand why I think this organism is a plant.

Here is an example of an essay written for the same prompt that is factually correct and effective at using information.

> I think that this creature should be in the Fungi Kingdom. I think this because of it is a decomposer. The organism has many cells. The organism reproduces by spore cases. The reason I would not put the organism in the Animal Kindom is because of the way it also uses photothintesis for food. I would not put it in the Moneran kingdom because it is one-celled. same with the Protist Kindom. I would not put it in the plant kingdom because the organism is a decomposer. That is why I would put in the Fungi kindom. I would name this creature a spordoprococellis. Because of the way it is a consumer, has many cells, producer, decomposer, and the way it reproduces-Sporangia.

Recall that validity has to do with the validity of inferences about a student's performance, not the validity of an assessment task or scoring procedure. What valid inferences could you make about the two students who wrote these essays?

Clearly both have used information effectively, but the second seems to have a better grasp of the factual content necessary for manipulating information. On a score that attends to the factual correctness of students' performance, the second student would receive a higher score. Is fact learning the same construct as using information?

Score of 4

This essay receives a score of "4" because it both sides of the argument in support of the choice the student made (the organism is a fungus). However, because only one argument was given for not placing the organism in other categories, the argument is not as robust as the two shown above that each received a score of "5."

> I think the fungi kigndom. Mainly because the majority of the characteristics are from this kingdom. Some features that make it a fungus are; many celled organism, can get food by breaking down waste products and dead tussues, and during reproduction it produces spores cases. For all of these reasons it makes it a member of the fungi kingdom. The main reasons it can't be placed in another are no other kingdoms have an organism with the amount of characteristics that the fungi kingdom does. Thus I conclude that the organism must be a member of the fungi family.

Score of 3

The following essay has all the attributes of a "3" essay. It makes a clear choice, and supports the choice with rationale. However, it fails to present both sides of the argument. We know why the author made up a new kingdom but we don't know why the organism doesn't fit into the existing kingdoms.

> I think this organism is a mix between plant, fungi, and animal kingdom. So I don't know what group to put it in. So I made up my own kingdom to put it in. Its call the mix-em-up kingdom. It should be classified tht way because this organism it reproduces with spore cases like a fungi. It makes it's own food by energy from the sun like a plant. And has many cells and eat other organisms like a animal. Thats my reson I put it in that kingdom.

Here is another way an essay can obtain a score of "3." This essay makes a clear choice (the organism is a fungus). Furthermore, the author presents multiple arguments for why the organism can't be classified in other kingdoms. However, only one reason is given for why the organism should be classified as a fungus. If more than one argument had been presented in favor of classifying it as a fungus, the essay would have received a higher score.

> I think it is a fungi. It can't be a protist or moneran because it's many celled. It can't be an animal because it produces food. It can't be a plant because plants aren't decompossers or consumers. It's a fungi because it produces spore cases.

Score of 2

The following essay makes a clear choice but only presents one rationale. It fails to present reasons for and against alternative choices and does not provide more than one reason for the choice that is made. This essay barely moves beyond the realm of opinion, although the statement that no existing kingdom fits all of the attributes of the organism is factually correct and if we give the student the benefit of the doubt, probably reflects learning that occurred as a result of recent instruction.

> I think we should make up a kingdom of our own. There is no kingdom to fit all the attributes of this organism. We should at least give it a name. We can't give it a kingdom if it does not have a name.

Here is another example of a "2" essay. The student did not make a clear choice. It is unclear how the author would classify the organism. Most of the essay supports the classification of the organism as a decomposer. Unfortunately, the first sentence adds a dimension of ambiguity that makes it difficult to understand the author's position.

> In the sumer this creature is a producer. Because of all the sun the island gets, and in the fall beacauss of the harsh winter they have all the trees and some animals it becomes a decomeposer so it formes on the things that have died or are roting and breakes them down and decomeposes it. then by the time the spring comes the creature is tirered of decomposing so he formes into a blob and he rolles throw the forest and eats-up all the paysness or harmfull orgaisims or plants. so I geus you could call the creature the maid of the island.

Score of 1

Here is an example of an essay that makes a clear choice but completely fails to present a rationale for the choice or against other options. We know the student classified the organism as a plant but we don't know why (within the context of the prompt).

> When a sciencetist was walking through the woods and found this weird plant. He walked back to the lab and looked it up he found that it was a Vienus fly trap. He put it in the plant family. And put it in the newspaper.

Here is another example of a "1" essay. This author makes no clear choice. This essay is minimally acceptable in the context of the prompt. The statements that the organism has many cells and has characteristics of multiple kingdoms are factually correct and could be used in defense of a choice if one were presented. However, this is all information that was presented in the prompt so it is difficult to infer that the student has actually used their knowledge of classifying organisms with a dichotomous key to solve the problem.

> It can be any one Because it has cahartersticks of all 3 of them and it has many cells.