CHAPTER 3

Summative Assessment

Chapter 2 has introduced various techniques for assessing student preconceptions. Following the Understanding by Design (UbD) approach (Wiggins & McTighe, 2005), the next stage in planning a unit of instruction is deciding what evidence at the end of the unit indicates student mastery of learning outcomes, which calls for the development of summative assessment. Planning summative assessment before planning learning activities is what makes the UbD's backward design approach unique and powerful because knowledge of what students should demonstrate at the end of the unit can inform what you should plan for students during the unit (i.e., opportunity to learn).

Summative assessment takes place at the end of a learning sequence to find out if students have mastered the learning outcomes. There are two important uses of summative assessment results in science teaching: (a) to decide if remedial instruction is necessary for some students to help them meet the learning expectations and (b) to grade students to create a record of their achievement and to communicate it to parents and interested others. Because the above two uses have serious implications in terms of allocation of learning resources, time, and future learning opportunities, summative assessment needs to be conducted with great care and thought. In this chapter, we will first develop skills in creating a test blueprint for summative assessment. We will then develop specific skills in writing high-quality multiple-choice, matching, and constructed-response questions and in providing differentiated assessment to students for their individual differences. We will also develop skills in reviewing and selecting external test questions for summative assessment.

CREATING TEST GRIDS

A very important first step in developing a summative assessment is to define the domain of assessment. The domain of assessment represents the scope and depth of test coverage; it is directly derived from the defined goals and objectives for the unit. One way to define the domain of assessment is to use a two-dimensional table commonly called a test grid or test blueprint. **A test grid** consists of a topic dimension and a cognitive reasoning dimension. Because there may be different emphases given to different topics and different cognitive

Essential Skills Addressed in Chapter 3

- Create a summative assessment test grid.
- Write high-quality multiple-choice questions.
- Write high-quality matching questions.
- Write high-quality constructed-response questions.
- Conducted differentiated assessment.
- Evaluate external test questions.

reasoning skills, it is necessary to assign different weights to different combinations of topics and cognitive reasoning skill. The above three components—that is, topic, cognitive reasoning skills, and weight—form a test grid with values. In a test grid with values, topics are usually in rows and cognitive reasoning skills in columns. Table 3.1 shows a sample test grid with values.

A test grid with values like Table 3.1 indicates two important aspects about the assessment domain: (a) what to assess (the intersections between rows and columns) and (b) how much emphasis there is for each combination in the assessment (the cell values). In the example of Table 3.1, we see that the assessment will cover five topics: substances, mixture, conservation, physical change, and chemical change; the above topics of content involve three cognitive reasoning skills: remembering, understanding, and applying. The cell values are determined by the product between weights of the corresponding topic and skill and the total points of the summative assessment. The procedure for creating such a table is as follows: (a) Identify topics to be assessed and their relative emphases in percentages (defining the rows), (b) identify cognitive skills expected and their relative emphases in percentages (defining the columns), (c) decide the total points for the summative assessment,

TABLE 3.1 A Sample Test Grid With Values

Values		*Remembering* *50%*	*Understanding* *30%*	*Applying* *20%*	*Subtotal (Points)*
Substances	*20%*	5	3	2	10
Mixture	*30%*	8	4	3	15
Conservation	*15%*	4	2	2	8
Physical change	*20%*	5	3	2	10
Chemical change	*15%*	4	2	1	7
Subtotal (points)		26	14	10	**50 points**

Assessment Standards Addressed in Chapter 3

NSES Assessment Standard A

Assessment must be consistent with the decisions they are designed to inform. This standard is further elaborated into the following substandards:

- Assessments are deliberately designed.
- Assessments have explicitly stated purposes.
- The relation between the decisions and the data is clear.
- Assessment procedures are internally consistent. (National Research Council [NRC], 1996, p. 78)

NSES Assessment Standard C

The technical quality of the data collected is well matched to the decisions and actions taken on the basis of their interpretation. This standard is further elaborated into the following substandards:

- The feature that is claimed to be measured is actually measured.
- An individual student's performance is similar on two or more tasks that claim to measure the same aspect of student achievement.
- Students have adequate opportunity to demonstrate their achievements.
- Assessment tasks and methods for presenting them provide data that are sufficiently stable to lead to the same decisions if used at different times. (NRC, 1996, pp. 83–85)

NSES Assessment Standard D

Assessment practices must be fair. This standard is further elaborated into the following substandards:

- Assessment tasks must be reviewed for the use of stereotypes, for assumptions that reflect the perspectives or experiences of a particular group, for language that might be offensive to a particular group, and for other features that might distract students from the intended tasks.
- Large-scale assessments must use statistical techniques to identify potential bias among subgroups.
- Assessment tasks must be appropriately modified to accommodate the needs of students with physical disabilities, learning disabilities, or limited English proficiency.
- Assessment tasks must be set in a variety of contexts, be engaging to students with different interests and experiences, and must not assume the perspective or experience of a particular gender, racial, or ethnic group. (NRC, 1996, pp. 85–86)

(d) calculate the cell values by multiplying total points by combined relative emphases, and (e) make adjustment to cell values and ensure that the total of cell values is equal to the total summative assessment points. Deciding emphases of topics and cognitive reasoning is subjective; important factors determining relative emphases may include importance in the curriculum, amount of instructional time to spend, and students' ability level. The total summative assessment point is arbitrary and decided mainly for convenience.

Once the assessment domain is defined in the form of a test grid with values, the next step to plan the summative assessment is to decide the assessment format and question type. There is a wide variety of assessment formats and question types from which to choose. The most commonly used assessment formats are paper-and-pencil tests and performance assessments. Paper-and-pencil tests are good at assessing knowledge and simple understanding, while performance assessments are good at assessing science inquiry. Both formats of assessment are typically necessary because good understanding needs to be demonstrated in multiple ways and contexts. Question types can be many, such as multiple-choice, matching, short constructed-response, extended constructed-response, and essay questions.

Once you have decided on assessment formats and question types, you can then operationalize the test grid with values into a test grid with items. Table 3.2 shows a sample test grid with items based on the test grid in Table 3.1.

From Table 3.2, we see that the summative assessment will include two tests (i.e., paper-and-pencil and performance tests). The paper-and-pencil test will include multiple-choice questions for assessing remembering and constructed-response questions for assessing understanding. The performance assessment will test students' ability to apply their knowledge of all the topics to conduct two performance tasks. We also see from Table 3.2 that the summative assessment will include 35 questions, among which 26 are multiple-choice questions, 7 constructed-response questions, and 2 performance tasks. The distribution of the questions is also indicated in the cells of the table. Relating cells of Table 3.2 to those of Table 3.1, we see that each multiple-choice question will have 1 point, each constructed-response question will have 2 or 3 points, and each performance assessment will have more than 1 point (such as 5 points for each task).

Once a test grid with items is created, the next steps in developing a summative assessment are to write test questions and develop performance assessment by using the test grid

TABLE 3.2 A Sample Test Grid With Items

Number of Items (Points)	*Remembering*	*Understanding*	*Applying*	*Subtotal*
	Multiple Choice	*Constructed Response*	*Performance*	
Substances	5 (5)	2 (3)	2 (10)	
Mixture	8 (8)	2 (4)		
Conservation	4 (4)	1 (2)		
Physical change	5 (5)	1 (3)		
Chemical change	4 (4)	1 (2)		
Subtotal	26 (26)	7 (14)	2 (10)	**35 (50)**

with items as a guide. This chapter will focus on paper-and-pencil test questions; Chapter 4 will focus on other types of test questions. But before we proceed, let's take a moment to reflect and apply what we have learned.

Application and Self-Reflection 3.1

An intermediate science teacher is planning a unit on energy transfer. The unit will take 4 weeks. The overall goal of the unit is to develop an overall understanding that energy exists in many different forms, and although energy can be transferred from one form to another, the total amount of energy is conserved. The overall objective of the unit is to help students describe the sources and identify the transformations of energy observed in everyday life. The specific objectives for the unit are as follows:

By the end of the unit, students will understand the following:

1. The sun is a major source of energy for the earth. Other sources of energy include nuclear and geothermal energy.

2. Fossil fuels contain stored solar energy and are considered nonrenewable resources. They are a major source of energy in the United States. Solar energy, wind, moving water, and biomass are some examples of renewable energy resources.

3. Most activities in everyday life involve one form of energy being transformed into another. For example, the chemical energy in gasoline is transformed into mechanical energy in an automobile engine. Energy, in the form of heat, is almost always one of the products of energy transformations.

4. Different forms of energy include heat, light, electrical, mechanical, sound, nuclear, and chemical. Energy is transformed in many ways.

5. Energy can be considered to be either kinetic energy, which is the energy of motion, or potential energy, which depends on relative position.

Develop a test grid with values and a test grid with items for a summative assessment of this unit. Share your test grids with the class and critique each other's.

WRITING MULTIPLE-CHOICE QUESTIONS

Multiple-choice (MC) questions are probably the most commonly used question type in science assessment. MC questions consist of two components: the stem and the choices. The stem is a statement asking a question, and the choices are possible answers, usually four or five, to the stated question. Choices must contain one best or correct answer; other choices are distracters or foils.

Guidelines for Writing Multiple-Choice Questions

1. The stem should be meaningful by itself and present a definite question.

Rationale: Presenting a question in the stem makes students understand the learning outcome you intend to assess. Otherwise, students have to read both the stem and all the choices to figure out what the question is asking for, which makes the question ambiguous and more likely for students to misunderstand your intention.

EXAMPLE The item should be meaningful

Poor A scientist . . .

a. Consults the writing of Aristotle.
b. Debates with fellow scientists.
c. Makes a careful observation during experiments.
d. Thinks about the probability.

Better How does a scientist typically discover new facts?

a. Consulting the writing of Aristotle.
b. Debating with fellow scientists.
c. Making careful observations during experiments.*
d. Thinking about the probability.

2. Items should be clear and in simple language.

Rationale: Learning outcomes in science are distinct from that in reading. It is necessary to separate reading and language skills from science achievements to increase the relevance of test scores.

EXAMPLE The item should be clear and in simple language

Poor In which state do records indicate a maximum statistical occurrence of earthquakes?

a. California*
b. Iowa
c. New York
d. South Carolina

Better Which state has most earthquakes on record?

a. California*
b. Iowa
c. New York
d. South Carolina

3. Make all choices plausible to uninformed students.

Rationale: Student understanding should be the only factor in determining whether the student will answer the question correctly. Implausible choices reduce the number of functional choices, thus increasing the probability for an uninformed student to answer the question correctly by guessing.

EXAMPLE Make choices plausible

Poor What are electrons?

a. Mechanical tools
b. Negative particles*
c. Neutral particles
d. Nuclei of atoms

Better What are electrons?

a. Negative particles*
b. Neutral particles
c. Nuclei of atoms
d. Positive particles

There are a few ways to make choices plausible to uninformed students. One way is to use typical student misconceptions identified during teaching or from published research. For example, research has shown that students commonly have the following misconceptions about motion: (a) All objects fall but heavy objects fall faster; (b) constant motion requires a constant force; (c) if a body is not moving, there is no force on it; and (d) if a body is moving, there is a force acting on it in the direction of motion. Accordingly, choices based on those misconceptions for an item on forces can be plausible for those uninformed students on mechanics.

4. Arrange the responses in an apparent logical order.

Rationale: Logically ordered choices suggest to students that there is no intention to place the correct answer in a particular position, and thus they should not speculate on the correct answer based on the order of choices. Because students will respond to the question based only on their understanding, the question will more likely serve its intended purpose.

EXAMPLE

Poor How many seasons in a year?

a. 3
b. 2
c. 4*
d. 1

(Continued)

(Continued)

Better How many seasons in a year?

a. 1
b. 2
c. 3
d. 4*

5. Avoid extraneous clues to the correct choice.

Rationale: Unnecessary information included in the choices may mislead a student to an incorrect answer or lead a student to the correct answer even though the student may not possess the required understanding. Extraneous clues can be a mismatch in grammar between the stem and the choice, suggesting the choice to be incorrect; unequal length in choices, misleading students to think the longer choice to be correct; or adverbs such as *absolutely* and *always* that are normally not correct.

EXAMPLE

Poor How do plants make their food?

a. Fertilizer
b. Photosynthesis*
c. Seed
d. Soil

Better What is necessary for plants to make their food?

a. Fertilizer
b. Photosynthesis*
c. Seed
d. Soil

6. Avoid using "none of the above" and "all of the above" alternatives.

Rationale: The use of "none of the above" and "all of the above" choices is usually due to exhaustion of adequate distracters or uncertainty on what the assessment objective of the item is. They are either redundant or misleading and thus need to be replaced by plausible choices. Even when "none of the above" or "all of the above" is the intended correct answer, the use of them is still unnecessary because these choices do not indicate exactly what students know or don't know.

EXAMPLE

Poor According to Boyle's law, which of the following changes will occur to the pressure of a gas at a given temperature when the volume of the gas is increased?

a. Increase
b. Decrease*
c. No change
d. None of the above

Better According to Boyle's law, which of the following changes will occur to the pressure of a gas at a given temperature when the volume of the gas is increased?

a. Increase
b. Decrease*
c. Increase first, then decrease
d. Decrease first, then increase
e. No change

EXAMPLE

Poor Which of the following is a fossil fuel?

a. Coal
b. Natural gas
c. Oil
d. All of the above*

Better Which of the following is NOT a fossil fuel?

a. Coal
b. Hydro*
c. Natural gas
d. Oil

7. Avoid using the "I don't know" choice.

Rationale: One good intention for the "I don't know" choice may be to discourage guessing. However, guessing is a reality with multiple-choice questions; if guessing is a concern, then other question types are available. The "I don't know" choice does not provide teachers or researchers any information on what a student knows or doesn't know. Some students who indeed don't know may select the "I don't know" choice and rightfully expect it to be the correct answer.

EXAMPLE

Poor Which of the following is a mammal?
a. Mosquito
b. Rat*
c. Spider
d. I don't know

Better Which of the following is a mammal?
a. Mosquito
b. Rat*
c. Spider

Techniques for Writing Multiple-Choice Questions for Assessing Higher Order Thinking

One common criticism of multiple-choice questions is that they assess lower order thinking (LOT). This criticism is not without merits. Given the details multiple-choice questions attend to, they are best at assessing factual knowledge, concrete reasoning, and discrete skills—LOT. However, the above do not necessarily mean that multiple-choice questions cannot assess higher order thinking (HOT). Using multiple-choice questions to assess HOT requires much more thoughts and skills. Before we discuss some techniques for assessing HOT using multiple-choice questions, let's first clarify what is LOT and what is HOT. LOT usually refers to the first three levels of the revised Bloom's cognitive taxonomy (i.e., remembering, understanding, and applying), and HOT usually refers to the last three levels of the revised Bloom's taxonomy (i.e., analyzing, evaluating, and creating). The common operational verbs for each of the above cognitive levels are as follows:

- Remember: recognize (identify), recall (retrieve)
- Understand: interpret (clarify, paraphrase, represent, translate), exemplify (illustrate, instantiate), classify (categorize, instantiate), summarize (abstract, generalize), infer (conclude, extrapolate, interpolate, predict), compare (contrast, map, match), explain (construct, model)
- Apply: execute (carry out), implement (use)
- Analyze: differentiate (discriminate, distinguish, focus, select), organize (find coherence, integrate, outline, parse, structure), attribute (deconstruct)
- Evaluate: check (coordinate, detect, monitor, test), critique (judge)
- Create: generate (hypothesize), plan (design), produce (construct)

Keeping in mind the above operational verbs, we now discuss some techniques for assessing HOT using multiple-choice questions.

Technique 1: Use Combinations of Question Formats

Example 1: MC + MC

Understand	*Analyze*
After a large ice cube has melted in a beaker of water, how will the water level change? a. Higher b. Lower c. The same*	After a large ice cube has melted in a beaker of water, how will the water level change? a. Higher b. Lower c. The same* Why do you think so? Choose all that apply. a. The mass of water displaced is equal to the mass of the ice.* b. Ice has more volume than water. c. Water is denser than ice. d. Ice cube decreases the temperature of water. e. Water molecules in water occupy more space than in ice.

Example 2: MC + Constructed Response

Understand	*Analyze*
After a large ice cube has melted in a beaker of water, how will the water level change? a. Higher b. Lower c. The same*	After a large ice cube has melted in a beaker of water, how will the water level change? a. Higher b. Lower c. The same* Why do you think so? Please justify your choice:

Technique 2: Provide a Factual Statement and Ask Students to Analyze

Example

Analyze
The sun is the only body in our solar system that gives off large amounts of light and heat. Select from the following the best reason for which we can see the moon. a. It is nearer to the earth than the sun. b. It is reflecting light from the sun.* c. It is the biggest object in the solar system. d. It is without an atmosphere.

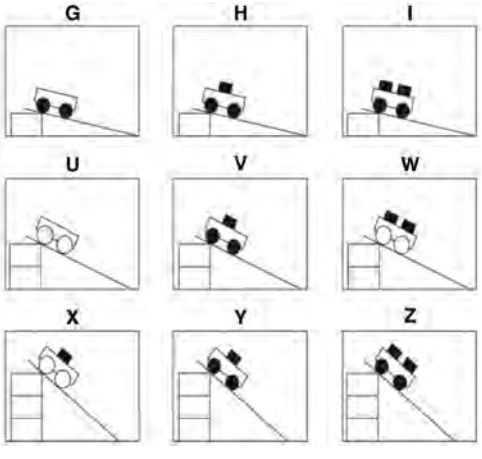
Technique 3: Provide a Diagram and Ask Students to Identify Elements

EXAMPLE

Analyze

The diagrams show nine different trials Usman carried out using carts with wheels of two different sizes and different numbers of blocks of equal mass. He used the same ramp for all trials, starting the carts from different heights. He wants to test this idea: The higher the ramp is placed, the faster the cart will travel at the bottom of the ramp. Which three trials should he compare?

a. G, H, and I
b. I, W, and Z
c. I, V, and X
d. U, W, and X
e. H, V, and Y*



Source: © International Association for the Evaluation of Educational Achievement (IEA), TIMSS 2003 Released http://timss.bc.edu/. Reproduced by permission.

Technique 4: Provide Data and Ask Students to Develop a Hypothesis

EXAMPLE

Create

Amounts of oxygen produced in a pond at different depths are shown below:

Location	Oxygen
Top meter	4 g/m^3
Second meter	3 g/m^3
Third meter	1 g/m^3
Bottom meter	0 g/m^3

Which statement is a reasonable hypothesis based on the data in the table?

a. More oxygen production occurs near the surface because there is more light there.*
b. More oxygen production occurs near the bottom because there are more plants there.
c. The greater the water pressure, the more oxygen production occurs.
d. The rate of oxygen production is not related to depth.

Technique 5: Provide a Statement and Ask Students to Evaluate Its Validity

EXAMPLE

Evaluate

You can hear a bell ringing in a vacuum chamber. How valid is this statement?

a. Valid
b. Partially valid
c. Invalid*
d. Not enough information to make a judgment

APPLICATION AND SELF-REFLECTION 3.2

Now let's try to practice what we have learned. For a topic on energy source (fifth grade), write one multiple-choice question for assessing each of the following cognitive reasoning skills: remembering, understanding, and evaluating. Present your questions to the class and evaluate each other's questions using the above guidelines.

WRITING MATCHING QUESTIONS

Matching questions typically assess students' understanding about relationships between concepts. A matching question consists of three components: (a) The *direction* poses a general problem, orients students to the question format, and instructs them on how to answer the question; (b) the *premises* elaborate on the general problem in the direction and pose a set of implicit questions; and (c) the *responses* provide potential answers as choices to the set of questions implied in the premises. Therefore, matching questions and MC questions share commonalities in requiring students to differentiate among choices and selecting the best ones to be matched. Different from MC questions, matching questions present a number of smaller questions of a general problem. The following is a sample matching question:

EXAMPLE

Different products involve different forms of energy. On the line at the left of each product, write the number of the energy form(s) from column B that the product involves. Energy forms in column B may be used once, more than once, or not at all.

(Continued)

(Continued)

	Column A	Column B
_____	a. Electric lightbulb	1. Electric energy
_____	b. Electric microscope	2. Light energy
_____	c. Gasoline engine	3. Magnetic energy
_____	d. TV	4. Mechanic energy
		5. Nuclear energy
		6. Potential energy
		7. Sound energy

As can be seen from the above example, a matching question covers a wide range of aspects of a topic. The two lists, columns A and B, are uneven in number, which reduces the chance of guessing by using one-to-one correspondence. Instructing students to choose items from column B once or more than once further reduces the chance of guessing. Finally, the elements in each of the two columns are homogeneous, meaning that they belong to a same category. Making both columns homogeneous helps define the specific learning outcome the matching question intends to assess. Other features are also helpful. For example, we can see from the above example that elements in both columns are arranged in a logical order (i.e., alphabetically), reducing unnecessary speculation on correct answers. Column A is about products, and column B is about energy transfer. Column B logically follows column A. In summary, essential characteristics of matching questions are as follows: (a) Use only homogeneous materials in both the premises and responses; (b) include an unequal number of responses and premises; (c) instruct students that responses may be used once, more than once, or not at all; and (d) arrange the list of premises and responses in a logical order.

Compared with MC questions, however, matching questions have limitations. The most obvious one is that matching questions are less focused in the content to be assessed. Second, matching questions allow students to make multiple one-to-one connections, thus creating a high potential for examinees to make guesses, although the probability of answering the entire question correctly by guessing remains low.

WRITING CONSTRUCTED-RESPONSE QUESTIONS

Compared with selected-response questions, constructed-response questions are open-ended in that students have to generate words or sentences to answer questions; there is no chance for them to answer questions correctly by random guessing. However, because students' responses are not constrained, a same constructed-response question may solicit quite different responses. Thus, constructed-response questions are more appropriate for assessing learning outcomes that are divergent and more cognitively advanced, beyond knowledge and comprehension levels of the Bloom's taxonomy.

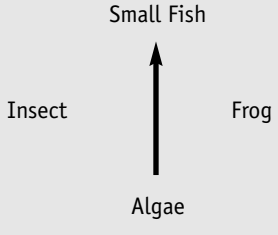
Depending on the extent of open-endedness, constructed-response questions can be (a) short constructed-response questions or (b) extended constructed-response questions.

Short Constructed-Response Questions

Short constructed-response questions require answers ranging from one word to a few sentences. Often, short constructed-response questions appear in a cluster following a brief introduction of a general problem, aspect, topic, concept, situation, and so on. Thus, like matching questions, a cluster of short constructed-response questions can cover a number of specific aspects of a concept or topic. Let's look at one example.

EXAMPLE

You will now finish a diagram of a food web in the pond. The food web shows what eats what in the pond system. Draw arrows in the diagram below from each living thing to the things that eat it. (The first arrow is drawn for you.)



Scoring key:

Complete: Student demonstrates a thorough understanding of the relationships in the food web by drawing four correct arrows and no incorrect arrows. Credited responses include:

a. Frog eats insect—arrow from insect to frog (1)
b. Frog eats algae—arrow from algae to frog (a juvenile frog eats algae) (1)
c. Insect eats algae—arrow from algae to insect (1)
d. Small fish eats insect—arrow from insect to small fish (1)

Partial: Student demonstrates some understanding of the relationships in the food web by drawing one, two, or three correct arrows and no incorrect arrows.

Unsatisfactory/Incorrect: Student demonstrates little or no understanding of the relationships in the food web by drawing all incorrect arrows, or both correct and incorrect arrows.

Source: National Assessment of Education Progress (http://nces.ed.gov/nationsreportcard/itmrls/).

The number in a bracket is the credit for correctly answering the question. As we can see, short constructed-response questions are very specific. Although no choices are

given, choices for correct answers are limited (often only one correct). Developing a complete and clear scoring rubric is critical to ensure that the questions assess the intended learning outcomes.

Extended Constructed-Response Questions

Compared with short constructed-response questions, extended constructed-response questions require more elaborate responses ranging from a few sentences to a short paragraph.

EXAMPLE Extended constructed-response questions should be more elaborate

Air is colorless, odorless, and tasteless. Describe one way that air can be shown to exist.

Scoring rubric
Correct Response:

1. Mentions that you can feel or see effects of air movement.
Examples: Wind, flags blowing, waving arms, and spreading smell.
2. Mentions that (light) things may float in air.
Example: A piece of paper floats in air.
3. Refers to the fact that air can be weighed.
4. Mentions that balloons, tires, and so on can be filled with air.
5. Refers to air pressure.
Example: Barometers show that air exists.
6. Refers to being able to "see" air.
Example: You can see air bubbles in water.
7. Other correct.

Incorrect Response:

1. We can breathe air. Refers only to the need of oxygen or air for life and other processes.
Examples: All living things need air/oxygen; candles extinguish without air; we would die if there was no air.
2. Refers to seeing water vapor.
Example: You can see water "vapor" when breathing out (on cold days or on a mirror or glass).
3. Merely repeats information in the stem.
4. Other incorrect.

Source: © International Association for the Evaluation of Educational Achievement (IEA), TIMSS 1995 Released http://timss.bc.edu/. Reproduced by permission.

As can be seen from the above example, an extended constructed-response question can be very brief, as short as one sentence. Because the expected answers are open-ended, the scoring scheme needs to be very detailed and comprehensive. Anticipation for all possible correct answers and incorrect answers is necessary for developing a good scoring rubric. In addition, it is a good idea to leave room for additional unexpected correct and incorrect answers.

Guidelines for Writing Constructed-Response Questions

1. Define the task completely and specifically.

EXAMPLE Define the task completely and specifically

Poor	Describe whether you think pesticides should be used on farms.
Better	Describe the environmental effects of pesticide use on farms.

2. Give explicit directions such as the length, grading guideline, and time frame to complete.

EXAMPLE Give explicit directions

Poor	State whether you think pesticide should be used on farms.
Better	State whether you think pesticide should be used on farms. Defend your position as follows: a. Identify any positive benefits associated with pesticide use. b. Identify any negative effects associated with pesticide use. c. Compare positive benefits against negative effects. d. Suggest if better alternatives than pesticides are available. Your essay should be no more than two double-spaced pages. Two of the points will be used to evaluate the sentence structure, punctuation, and spelling (10 points).

3. Do not provide optional questions for students to choose.

This is because different questions may measure completely different learning outcomes, which makes comparisons among students difficult.

4. Define scoring clearly and appropriately.

A scoring rubric can be either analytic or holistic. Please refer to Chapter 4 for guidelines.

Application and Self-Reflection 3.3

Let's use the same topic—that is, energy sources (fifth grade). Write one short constructed-response question and one extended constructed-response question at the understanding cognitive reasoning level. Your questions should include scoring rubrics. Present your questions to the class and critique each other's questions.

DIFFERENTIATED SUMMATIVE ASSESSMENTS

One major challenge in today's science classroom is student individual differences. Given any class of students, it is easy to identify students with differences in their prior knowledge, motivation to learn, social-cultural values of science, and learning style, to name just a few. The National Science Education Standards (NRC, 1996) call for achieving science literacy by all students. While group differences in student achievements and opportunity to learn in terms of gender, ethnicity, socioeconomic status, and so on may require institutional efforts to address, student individual differences are an important part of classroom teachers' responsibilities. Effective science teachers teach students as individuals rather than as a group.

Student individual differences may exist in many ways. Postlethwaite (1993) classifies student individual differences into the following categories: (a) educational differences, (b) psychological differences, (c) physical differences, and (d) other differences. The educational differences may include knowledge, reasoning, reading, and problem solving. Psychological differences may include learning style, motivation, personality, locus of control, and IQ; physical differences may include differences related to the senses of vision, hearing, touching, and smelling. Other differences can be those related to socioeconomic status, gender, religious beliefs, and so on. Given the ubiquitous differences exhibited in any classroom, differentiated science instruction is a necessity.

Differentiated science instruction calls for differentiated assessment. **Differentiated assessment** is an approach to conducting assessment according to student individual differences. Differentiated assessment and differentiated instruction, although closely related, are distinct from each other based on their purposes. While differentiated instruction aims at providing ideal learning opportunities in order for every student to achieve his or her maximal learning potential, differentiated assessment aims at providing ideal conditions in order for every student to demonstrate his or her best science achievement.

Differentiated assessment has the following two basic principles:

Principle 1: Students should be assessed under conditions that best enable them to demonstrate their achievement.

Principle 2: Results from different assessment methods for different students should be related in order for a same scoring system to be used.

Principle 1 requires that assessment methods used match students' individual differences. For example, a more difficult and challenging test may be provided to those gifted

and talented students. Students with visual impairment may be assessed using large prints or Braille. Principle 2 requires that different assessment methods for different students must assess related learning outcomes, and the scores resulting from different assessment methods must be comparable. This principle is particularly important in today's standards-based approach to science teaching and learning in which students are expected to achieve a common set of learning standards. Principle 2 also enables a common grading system (to be discussed in Chapter 7) to be used no matter what assessment methods students may have taken. To make scores from different assessment methods comparable, you must have a common ground, such as a common set of questions among different assessment methods. Principles 1 and 2 complement each other; together they form a valid differentiated assessment system.

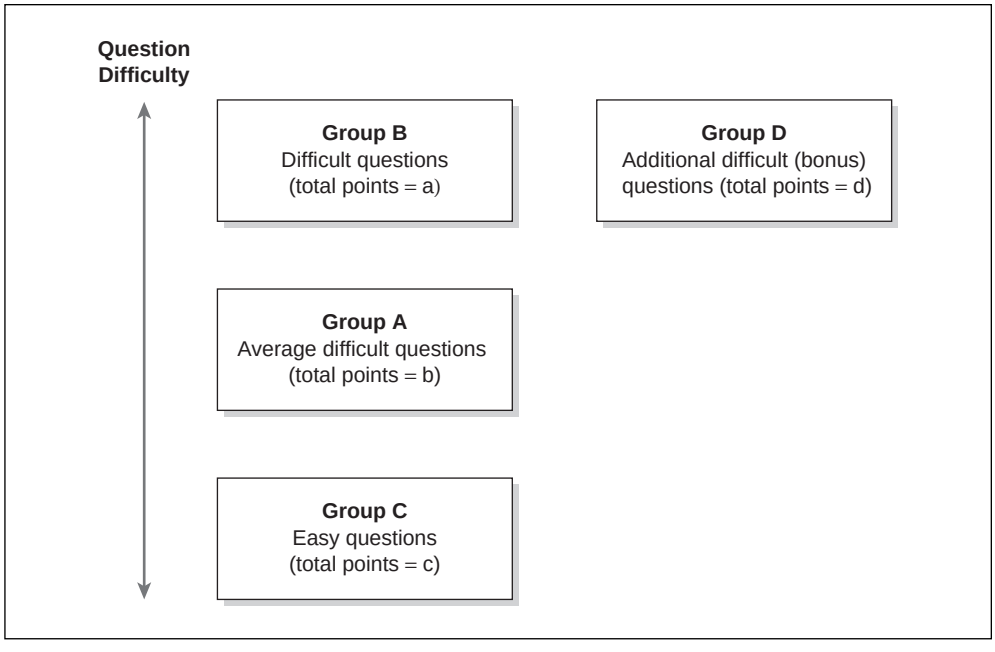
Next we will discuss some common differentiated assessment methods applicable to common student individual differences. More severe differences formally identified under the Individuals with Disabilities Education Act (IDEA), such as physical handicaps, learning disabilities, and attention-deficit hyperactivity disorder (ADHD), require assessment accommodation, adaption, or alternative tests, which will be discussed in Chapter 5.

Differentiated Assessment for Educational Differences

Although educational differences can be in various forms, they can be classified into two types: science related and nonscience related. Science-related individual differences refer to different levels of science achievement by students, with some more advanced and others less advanced. Psychometric theories suggest that most valid and reliable tests are the ones that match students' abilities. That is, more advanced students should take more difficult tests, while less advanced students should take easier tests. Therefore, differentiated assessment for students of different science achievement levels requires alternative forms with different difficulties. For non-science-related individual differences such as different reading levels and different writing abilities, test difficulty may remain the same, but the test questions may be presented at different reading or writing levels. In this case, alternative forms of the same difficulty are necessary.

Developing Alternative Forms of Different Difficulties

After you have developed questions based on your test grid with items, you will arrange your test questions from easiest to most difficult based on your interpretation of the curriculum standards and students' capability. Divide the test questions into three groups: easier questions, average difficult questions, and difficult questions. An ideal distribution would be ½ of total questions in the average group and ¼ of total questions each in the easier and more difficult groups. However, whatever distribution results, stay with it; you should not modify or drop questions to obtain a preferred distribution because your test questions have been developed according to your test grids—the defined assessment domain. Changing questions at this time will result in changing your assessment domain. If there are not enough difficult questions for more advanced students, you may add additional difficult questions as bonus questions for students to receive extra credit. The above grouping may be represented in the following diagram:



Once questions are grouped in the way described above, then three alternative forms of the same test may be created: Test Form A consists of Group A and a few Group B questions for average students; Test Form B consists of Group A and Group B questions, plus Group D (if available) questions for advanced students; and Test Form C consists of Group C and a few Group A questions for struggling students. The scoring of the above alternative forms is as follows:

Test Form A: total score = scores earned on questions of Groups A and B + c

Test Form B: total score = scores earned on questions of Groups A and B + c + bonus (scores earned on Group D questions)

Test Form C: total score = scores earned on questions of Groups A and C

Developing Alternative Forms of a Same Difficulty

Differentiated assessment using alternative forms of a same difficulty requires that the alternative forms assess the same learning outcomes, although the formats of test questions may be different. This can be the case when some questions of a same test are presented at different reading levels. Although reading and science achievement may be related, since the focus of a science course is to improve student science achievement, not reading ability, a required high reading level may inhibit students of lower reading ability to demonstrate their best science achievement. In this case, developing an alternative test form at a lower reading level is necessary.

To develop an alternative test form at a lower reading level, you need first of all to understand factors affecting readability. Many measures of readability are available; two of them are the Flesch Reading Ease Test (FRET) and the Flesch-Kincaid Grade Level Test (FKGLT). FRET is a scale from 0 to 100; the higher the score, the easier it is to read. A FRET score of 60 to 70 is about the seventh- or eighth-grade reading level. The FRET formula is as follows:

$$\text{FRET} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}),$$

where ASL = average sentence length (the number of words divided by the number of sentences), and ASW = average number of syllables per word (the number of syllables divided by the number of words).

Similar to FRET, FKGLT evaluates readability based on the average sentence length and average number of syllables per word. However, FKGLT produces a score that is equivalent to grade level. For example, a FKGLT score of 8 means that the reading level is about the eighth grade. FKGLT is calculated as follows:

$$\text{FKGLT} = (0.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.59,$$

where ASL and ASW are the same as in FRET.

MS Word 2007 produces both FRET and FKGLT scores automatically every time after you have completed a spelling check of your document. If you do not see a summary of readability statistics after the spelling check, you need to check the Show Readability Statistics option by taking the following procedures:

1. Click the **Microsoft Office Button**  , and then click **Word Options.**
2. Click **Proofing.**
3. Make sure **Check grammar with spelling** is selected.
4. Under **When correcting grammar in Word,** select the **Show readability statistics.**

If you use older versions of MS Word, the readability statistics is shown by selecting **Properties** under the pull-down menu of **File.**

From the above readability statistics, you can see that using shorter sentences and shorter words is the effective way to reduce the overall reading level of a test. If you already know the range of your students' approximate reading levels, you may decide to develop two alternative test forms, with one aiming at the lower end of the reading level range and another at the average level. For example, if your students' reading levels range from Grades 4 to 8, then you may develop one test form at Grade 4 level and another test form at Grade 6 level. MS Word's readability statistics will help you to make the necessary adjustment of reading levels.

Developing alternative forms of a test of the same difficulty for different writing abilities pertains only to extended constructed-response questions, such as short and long essays. For these questions, the alternatives can ask students to use drawing or other media forms such as audiotaping. Once again, no matter what medium is provided as an alternative for students, the question and scoring must remain the same to ensure that the same learning outcome is assessed.

Differentiated Assessment for Psychological Differences

Various psychological differences may exist in a classroom as well. One of the most important psychological differences commonly encountered by science teachers is learning style. **Learning style** refers to the preferred approaches to the acquisition and organization of knowledge. Simply put, learning style is about how students learn best. For example, some students learn best through hands-on experience, while others learn best through reading. Learning style is closely related to cognitive style. **Cognitive style** refers to the organization and control of cognitive processes pertaining to information receiving, contexts, and information processing. For example, field-dependent individuals may be highly sensitive to contextual factors such as visual cues when solving a problem, while field-independent individuals may rely more on personal knowledge and self-appreciation when solving a problem. As another example, Howard Gardner (1983) differentiated seven distinct intelligences people may exhibit: linguistic intelligence, musical intelligence, logical-mathematical intelligence, spatial intelligence, bodily-kinesthetic intelligence, interpersonal intelligence, and intrapersonal intelligence. Cognitive styles shape learning styles.

When thinking about student individual differences in terms of learning styles, it is important to understand that learning styles are process variables; they do not equate to outcomes. Research has shown that the correlation between learning styles (i.e., field dependence/independence and locus of control) and science achievement is from modest to low for declarative knowledge (Baker & Piburn, 1997). Thus, for declarative knowledge in science, "styles are not better or worse. They are just different" (Baker & Piburn, 1997, p. 253). Thus, for the assessment of declarative knowledge, a question in favor of a particular learning style may place students of other learning styles at a disadvantage. However, for other types of knowledge, such as procedure knowledge, there is high correlation between learning style (i.e., field dependence/independence) and achievement. Thus, assessment of those types of knowledge needs to make sure if a particular learning style is indeed part of the assessment domain. If it is, differentiated assessment would actually undermine the validity of the assessment results and thus is not justified. For example, assessment of students' skills in performing an acid-base titration requires students to attend to contextual factors of burette, calibration solution, sample solution, indicators, and so on, which may match students with a learning style associated with field dependence. Differentiated assessment using paper-and-pencil tests to assess the acid-base titration lab skill would not be able to assess the same learning outcome as a hands-on manipulative test to assess the acid-base titration lab skill. Differentiated assessment is not an automatic option; it must be justified based on the consideration that alternative forms of assessment do not alter the intended assessment domain.

Various categorizations of student learning styles are available. One common categorization is in terms of preferred modality, which classifies learning styles to be visual, auditory, and hepatic. A student may demonstrate primarily one type of learning style or a combination of two or three. Table 3.3 summarizes characteristics of each learning style and the possible assessment differentiations.

Another common categorization is based on the preferred way of what and how information is acquired and processed during learning. There are two dimensions with each consisting of two bipolar modes, concrete versus abstract and sequential versus random. The

TABLE 3.3 Differentiated Assessment Methods for Learning Styles Based on Preferred Modality

Learning Style	*Characteristics*	*Differentiated Assessment*
Visual	1. Needs to see in order to understand 2. Strong sense of color 3. Artistic	1. Use graphical questions 2. Allow drawing and color pens to answer questions 3. Use flowchart to give directions
Auditory	1. Needs to hear in order to understand 2. Strong sense of sound 3. Good at oral communication	1. Use tape recording for instruction 2. Oral administration of tests 3. Audiotaping of responses to questions 4. Oral presentation
Haptic	1. Learns best by hands-on 2. Can assemble parts without reading directions 3. Enjoys physical activities 4. Shows good physical coordination and good in athletics	1. Performance assessment 2. Role-playing 3. Demonstration of responses to questions 4. Model creation

combination of the four poles results in four typical learning styles: concrete sequential (CS), abstract random (AR), concrete random (CR), and abstract sequential (AS). A student may demonstrate primarily one or a combination of the learning styles. Table 3.4 summarizes differentiated assessment methods for each of the four typical types of learning styles.

APPLICATION AND SELF-REFLECTION 3.4

1. Assume that you are going to create a test based on the test grid in Table 3.4. Describe a possible plan to develop two forms, with one for most students and another form for a few who are inclusion students with a lower academic performance level.

2. A fifth-grade teacher has just finished a unit on plants. The teacher's resource book has an end-of-unit test that consists of entirely multiple-choice questions. Discuss how you are going to create an alternative test form of the test with the same difficulty by attending to the following individual differences:

   a. Reading level at Grade 3

   b. Visual learners

TABLE 3.4 Differentiated Assessment Methods for Learning Styles Based on Preferred Information Acquisition and Processing

Learning Style	*Characteristics*	*Differentiated Assessment*
Concrete sequential (CS)	1. Ordered and structured 2. Detail and accuracy oriented 3. Hands-on 4. Prefers practical work with detailed directions	1. Selected response questions 2. Performance assessment 3. Data analysis 4. Structured problem solving 5. Interpreting or creating organization charts
Concrete random (CR)	1. Independent 2. Creative 3. Risk taker 4. Experimenter 5. Problem solver	1. Extended constructed-response questions 2. Open-ended problem solving 3. Project/long-term investigation 4. Model creation
Abstract random (AR)	1. Sensitive 2. Emotional 3. Imaginative 4. Personal 5. Flexible	1. Group project 2. Essay 3. Artistic expression 4. Model creation
Abstract sequential (AS)	1. Logical 2. Structured 3. Deep thinker 4. Evaluative	1. Selected response questions 2. Details instructions 3. Essay 4. Research project 5. Library research

EVALUATING EXTERNAL TEST QUESTIONS

Writing quality test questions requires time and skill. On the other hand, many assessment resources are available. For example, teacher resources accompanying a textbook may contain end-of-unit tests or exercises. Many states may also put previous years' state standardized tests online. Incorporating external test questions to develop summative assessment can not only save teachers' valuable time but also potentially enhance the quality of your summative assessment.

However, external test questions are not necessarily of high quality. Critically evaluating them is necessary. We will only discuss evaluation of individual test questions in this section; we will deal with the issue of using an entire external test in Chapter 5.

The first consideration when evaluating an external question is its relevance to the learning outcomes of your summative assessment. The relevance refers to the fit into your assessment domain as defined by your summative assessment test grid. If a test question fits into one of the cells of the test grid, the question is potentially useful. Otherwise, no further

evaluation is needed. If a test question is relevant, its technical quality is the next focus of evaluation. The guidelines discussed above for writing various types of questions can be used as criteria to evaluate the quality of the assessment questions. For example, if the question is a multiple-choice question, the seven guidelines may be applied to evaluate the question. If a question meets all the guidelines, the question can be incorporated directly into your summative assessment. Often a question does not meet one or a few criteria. In this situation, you can either abandon the question or modify it to make it meet all the criteria.

The following checklist has been created to facilitate evaluation of test questions. Please note that true-false and fill-in-the-blank types of questions are not discussed in this chapter. This is because true-false questions involve too much guessing, and fill-in-the-blank questions encourage too much memorization. Since what true-false and fill-in-the-blank questions assess can also be assessed or even better assessed by multiple-choice questions, matching questions, or constructed-response questions, there is no need to use true-false and fill-in-the-blank question types.

Checklist for Evaluating External Test Questions

Multiple Choice		
1. The stem of an item is meaningful by itself and presents a definite question.	Yes	No
2. The item is written in clear and simple language.	Yes	No
3. All choices are plausible to uninformed students.	Yes	No
4. Choices are in a logical order.	Yes	No
5. There are no extraneous clues to the correct choice.	Yes	No
6. There are no "none of the above" and "all of the above" alternatives.	Yes	No
7. There is no "I don't know" choice.	Yes	No
8. The language including vocabulary is appropriate for the students.	Yes	No
Matching		
1. There are homogeneous elements in both premises and responses.	Yes	No
2. There are an unequal number of responses and premises.	Yes	No
3. Premises and responses are in a logical order.	Yes	No
4. Responses may be used once, more than once, or not at all.	Yes	No
5. The language including vocabulary is appropriate for the students.	Yes	No

(Continued)

(Continued)

Constructed Response		
1. The task is defined completely and specifically.	Yes	No
2. There are explicit directions such as the length, grading guideline, and time to complete.	Yes	No
3. There are no optional questions for students to choose.	Yes	No
4. Scoring is clear and appropriate.	Yes	No
5. The language including vocabulary is appropriate for the students.	Yes	No

Application and Self-Reflection 3.5

Locate some test questions from a teacher's resource book, a student textbook, or any other sources and evaluate the appropriateness of the test questions by using the above checklist.

The Cases of Eric and Elisia: Summative Assessment

Before beginning this chapter, Eric and Elisia thought themselves to be good at writing test questions. They took numerous tests from elementary school to university, and they knew what types of questions they would use to develop a test. The test grids introduced in this chapter opened their eyes about the test development process, and they began to appreciate that test questions by themselves may not be said to be good or bad; a more meaningful question to ask is whether those questions serve intended purposes. For end-of-unit summative assessment, a common purpose is to find out if students have mastered the stated unit objectives or learning standards. Eric and Elisia can see how test grids are valuable to them to ensure that each question they are going to write for the end-of-unit test serves a specific purpose. They feel that the guidelines for writing multiple-choice questions and constructed-response questions, although mostly common sense, can potentially help them write high-quality questions. However, using multiple-choice questions to assess higher order thinking skills is something they never thought of. For Eric, since he teaches elementary grades, he still thinks a paper-and-pencil test is of only limited use for him; he can see how differentiated assessment is important to his situation, given that the diversity among students in his class cannot be greater. For Elisia, given that state tests use primarily

multiple-choice and constructed-response questions, she can see how she will be able to develop better end-of-unit tests by following the guidelines introduced in this chapter. She also sees the value of developing alternative forms of summative assessment because she has always struggled with the issue of accommodating student individual differences. For example, a few advanced students always complain that her tests are too easy. Overall, both Eric and Elisia feel that this chapter offers some practical suggestions for them to create an end-of-unit test, but they also feel that more variety of assessment question formats is needed. They look forward to subsequent chapters for more assessment ideas.

Do the experiences of Eric and Elisia sound familiar to you? What were your initial ideas of summative assessment, and how have they changed as the result of this chapter?

Chapter Summary

- Quality summative assessment depends on a systematic plan that matches intended learning outcomes. A test grid is a two-dimensional table defining the assessment domain in terms of topics and cognitive skills.
- Multiple-choice questions should contain a stem that poses a meaningful question and equally plausible choices, including a clearly correct answer. Choices should be arranged in a logical order; avoid extraneous clues and such choices as "none of the above," "all of the above," and "I don't know."
- Matching questions should contain two uneven numbers of homogeneous columns that are arranged in a logical order.
- Short and extended constructed-response questions should require clear and specific answers that can be scored either correctly or incorrectly. The analytic scoring scheme should be specific and comprehensive enough to encompass correct, partially correct, and incorrect responses.
- There is a wide variety of individual differences among students in a class. Differentiated assessment provides students with alternative assessment forms appropriate for educational, psychological, and other differences. Alternative assessment forms may be developed to accommodate different levels of science achievement or different learning styles. Alternative assessment forms can be with a same difficulty or different difficulties with common questions.
- There are many external sources for test questions. Not every external question is of high quality. When evaluating external test questions, you should first decide if they fit into the assessment domain defined by a test grid. Next it is necessary to review if test questions are properly written.

√ Mastery Checklist

- ☐ Create a summative assessment test grid.
- ☐ Write high-quality multiple-choice questions.
- ☐ Write multiple-choice questions to assess higher order thinking skills.
- ☐ Write high-quality matching questions.
- ☐ Write high-quality constructed-response questions.
- ☐ Develop alternative forms of tests of different difficulties for different achievement levels of students.
- ☐ Develop alternative forms of tests of a same difficulty for different learning styles.
- ☐ Evaluate external test questions.

Web-Based Student Study Site

The Companion Web site for *Essentials of Science Classroom Assessment* can be found at **www.sagepub.com/liustudy.**

The site includes a variety of materials to enhance your understanding of the chapter content. Visit the study site to complete an online self-assessment of essential knowledge and skills introduced in this chapter. The study materials also include flash cards, Web resources, and more.

Further Reading

Gallagher, J. D. (1998). *Classroom assessment for teachers.* Upper Saddle River, NJ: Merrill.

Although not science specific, this popular assessment textbook contains chapters on developing all types of paper-and-pencil test questions and various alternative assessments such as portfolios and performance assessments. It also contains chapters related to planning for assessment, grading, and standardized testing.

References

Baker, D. R., & Piburn, M. D. (1997). *Constructing science in middle and secondary school classrooms.* Boston: Allyn & Bacon.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

National Research Council (NRC). (1996). *National science education standards.* Washington, DC: National Academy Press.

Postlethwaite, K. (1993). *Differentiated science teaching.* Buckingham, UK: Open University Press.

Wiggins, G., & McTighe, J. (2005). *Understanding by design.* Alexandria, VA: Association for Supervision and Curriculum Development.