



pennsylvania
DEPARTMENT OF EDUCATION

The Pennsylvania System of School Assessment

Science Item and Scoring Sampler



2015–2016
Grade 8

Pennsylvania Department of Education Bureau of Curriculum, Assessment, and Instruction—August 2015

TABLE OF CONTENTS

Introduction	1
What Is Included	1
Purposes and Uses	1
Item Format and Scoring Guidelines	1
Testing Time and Mode of Testing Delivery for the PSSA	1
Item and Scoring Sampler Format	2
Science Test Directions	3
General Description of Scoring Guidelines for Science Open-Ended Questions	4
Multiple-Choice Questions	5
Open-Ended Questions	20
Science Grade 8—Summary Data	33

INTRODUCTION

The Pennsylvania Department of Education provides districts and schools with tools to assist in delivering focused instructional programs aligned with the Pennsylvania Academic Standards. In addition to the Academic Standards, these tools include Assessment Anchor documents, assessment handbooks, and content-based item and scoring samplers. Each Item and Scoring Sampler is a useful tool for Pennsylvania educators in preparing local instructional programs and can also be useful in preparing students for the statewide assessment.

WHAT IS INCLUDED

This sampler contains test questions, or test “items,” that have been written to align to the Assessment Anchors that are based on the Pennsylvania Academic Standards (PAS). The sample test questions model the types of items that will appear on an operational PSSA. Each sample test question has been through a rigorous review process to ensure alignment with the Assessment Anchors prior to being piloted in an embedded field test within a PSSA assessment and then used operationally on a PSSA assessment. Answer keys, scoring guidelines, and any related stimulus material are also included. Additionally, sample student responses are provided with each open-ended item to demonstrate the range of responses that students provided in response to these items.

PURPOSES AND USES

The items in this sampler may be used as models for creating assessment items at the classroom level, and they may also be copied and used as part of a local instructional program.¹ Classroom teachers may find it beneficial to have students respond to the open-ended items in this sampler. Educators can then use the item’s scoring guideline and sample responses as a basic guide to score the responses, either independently or together with colleagues within a school or district. The sampler also includes the *General Description of Scoring Guidelines for Science Open-Ended Questions* used to develop the item-specific guidelines. The general description of scoring guidelines can be used if any additional item-specific scoring guidelines are created for use within local instructional programs.¹

ITEM FORMAT AND SCORING GUIDELINES

The multiple-choice (MC) items have four answer choices. Each correct response to an MC item is worth one point.

Each open-ended (OE) item in science is scored using an item-specific scoring guideline based on a 0–2 point scale.

TESTING TIME AND MODE OF TESTING DELIVERY FOR THE PSSA

The PSSA is delivered in traditional paper-and-pencil format as well as in an online format. The estimated time to respond to a test question is the same for both methods of test delivery. During an official testing administration, students are given additional time as necessary to complete the test questions. The following table shows the estimated response time for each item type.

Science Item Type	MC	OE
Estimated Response Time (in minutes)	1	5

¹ The permission to copy and/or use these materials does not extend to commercial purposes.

ITEM AND SCORING SAMPLER FORMAT

This sampler includes the test directions and scoring guidelines that appear in the PSSA Science assessments. Each sample multiple-choice item is followed by a table that includes the alignment, answer key, DOK, the percentage² of students who chose each answer option, and a brief answer option analysis or rationale. Each open-ended item is followed by a table that includes the item alignment, DOK, and the mean student score. Additionally, each of the included item-specific scoring guidelines is combined with sample student responses representing each score point to form a practical, item-specific scoring guide. The General Description of Scoring Guidelines for Science used to develop the item-specific scoring guidelines should be used if any additional item-specific scoring guidelines are created for use within local instructional programs.

Example Multiple-Choice Item Information Table

Item Information				Option Annotations	
Alignment		Assigned AA/EC		Brief answer option analysis or rationale	
Answer Key		Correct Answer			
Depth of Knowledge		Assigned DOK			
<i>p</i> -values					
A	B	C	D		
Percentage of students who selected each option					

Example Open-Ended Item Information Table

Alignment	Assigned AA/EC	Depth of Knowledge	Assigned DOK	Mean Score	
------------------	----------------	---------------------------	--------------	-------------------	--

² All *p*-value percentages listed in the item information tables have been rounded.

SCIENCE TEST DIRECTIONS

Below are the test directions available to students taking the paper-and-pencil version of the assessment. These directions may be used to help students navigate through the assessment.

Directions:

On the following pages are the Science questions. There are several types of questions.

Multiple-Choice Questions

Some questions will ask you to select an answer from among four choices. These questions will be found in your test booklet.

For the multiple-choice questions:

- Read each question, and choose the best answer.
- Record your choice in the answer booklet.
- Only one of the answers provided is the correct response.

Science Scenario Multiple-Choice Questions

Some of the questions will require you to use some information found in a science scenario. The science scenario may contain text, graphics, charts, and/or tables, and may use these elements to describe the results of a class project, an experiment, or other similar research. The science scenario may cover several pages in your test booklet, and it will include information you will need to answer several multiple-choice questions.

For the science scenario multiple-choice questions:

- Read each question, and choose the best answer.
- Record your choice in the answer booklet.
- Only one of the answers provided is the correct response.

Open-Ended Questions

Other questions will require you to write your response. These questions will be found in your answer booklet.

For the open-ended questions:

- Be sure to read the directions carefully.
- If the question asks you to do two tasks, be sure to complete both tasks.
- If the question asks you to compare, be sure to compare. Also, if the question asks you to explain, describe, or identify, be sure to explain, describe, or identify.

GENERAL DESCRIPTION OF SCORING GUIDELINES FOR SCIENCE OPEN-ENDED QUESTIONS

2 POINTS

- The response demonstrates a *thorough* understanding of the scientific content, concepts, and procedures required by the task(s).
- The response provides a clear, complete, and correct response as required by the task(s). The response may contain a minor blemish or omission in work or explanation that does not detract from demonstrating a *thorough* understanding.

1 POINT

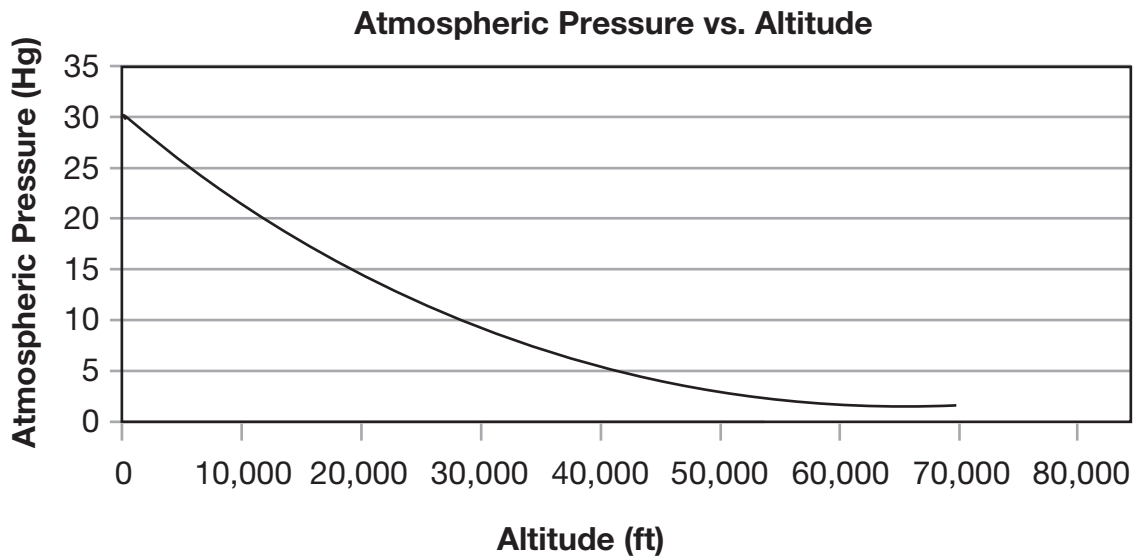
- The response demonstrates a *partial* understanding of the scientific content, concepts, and procedures required by the task(s).
- The response is somewhat correct with *partial* understanding of the required scientific content, concepts, and/or procedures demonstrated and/or explained. The response may contain some work that is incomplete or unclear.

0 POINTS

- The response provides *insufficient* evidence to demonstrate any understanding of the scientific content, concepts, and procedures as required by the task(s) for that grade level.
- The response may show only information copied or rephrased from the question or *insufficient* correct information to receive a score of 1.

MULTIPLE-CHOICE QUESTIONS

Use the graph below to answer question 1.



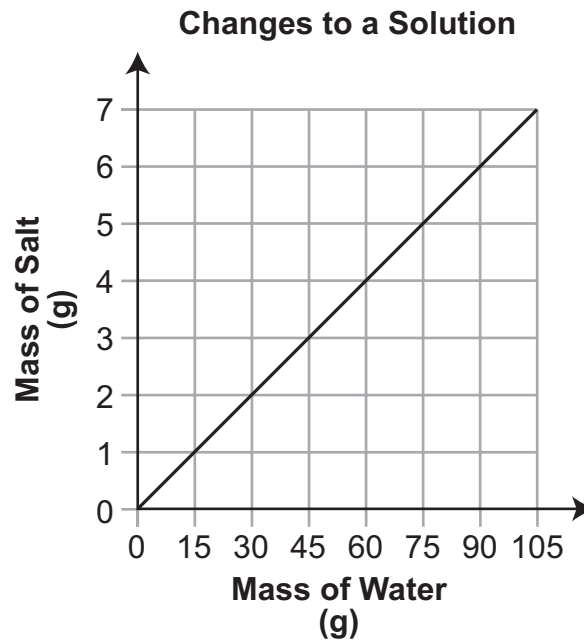
1. Based on the graph, which statement **best** describes the relationship between atmospheric pressure and altitude?
 - A. Atmospheric pressure controls altitude.
 - B. Atmospheric pressure is equal to altitude.
 - C. Atmospheric pressure increases as altitude increases.
 - D. Atmospheric pressure decreases as altitude increases.

Item Information				Option Annotations
Alignment		S8.A.1.1.3		A. Atmospheric pressure does not control altitude. B. Atmospheric pressure does not equal altitude. C. Atmospheric pressure decreases as altitude increases. D. Key: As altitude increases, atmospheric pressure decreases.
Answer Key		D		
Depth of Knowledge		2		
p-values				
A	B	C	D	
10%	4%	16%	70%	

2. In 1901, the SS *Port Morant* became the world's first refrigerated banana ship. It was equipped with carbon dioxide refrigeration. It could carry 23,000 bunches of bananas from Jamaica to England at a controlled temperature. What impact did cargo ship refrigeration systems have on the banana industry?
- A. It made transportation of bananas to overseas markets quicker.
 - B. It made bananas ripen by the time they arrived at their destination.
 - C. It made bananas sweeter than bananas transported without any cooling.
 - D. It made shipment of bananas to faraway ports possible with little spoilage.

Item Information				Option Annotations
Alignment		S8.A.1.2.4		A. Refrigeration does not affect the speed at which the ship moves. B. Refrigeration does not cause ripening. C. Refrigeration does not control the sugar content of bananas. D. Key: Refrigeration slows the rate at which food is spoiled.
Answer Key		D		
Depth of Knowledge		2		
p-values				
A	B	C	D	
14%	13%	7%	66%	

Use the graph below to answer question 3.



3. The graph shows the relationship between two characteristics of a saltwater solution. Which ratio describes the changes in these two characteristics?
- A. 1:7
 - B. 90:7
 - C. 1:15
 - D. 105:15

Item Information				Option Annotations
Alignment		S8.A.1.3.1		A. Based on the data, a 1:7 ratio of salt to water is too low. B. Based on the data, a 90:7 ratio of salt to water is too high. C. Key: 1 g of salt to 15 g of water is used to make the solution. D. Based on the data, a 105:15 ratio of salt to water is too high.
Answer Key		C		
Depth of Knowledge		2		
p-values				
A	B	C	D	
10%	9%	74%	7%	

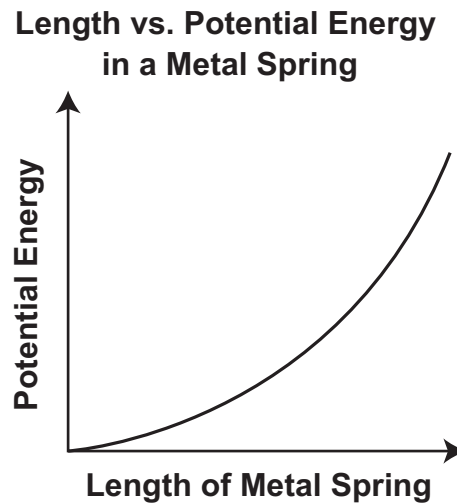
Use the data table below to answer question 4.

Trial	Time (seconds)		
	Mouse 1	Mouse 2	Mouse 3
1	58	52	67
2	54	50	65
3	53	49	61
4	47	48	57
5	42	46	55

4. A researcher placed three mice in a maze and recorded the time it took each mouse to complete the maze. Which relationship is **best** supported by the data collected by the researcher?
- A. More practice resulted in faster maze completion times for each mouse.
- B. More practice had no effect on the maze completion times for each mouse.
- C. More practice had the greatest effect on Mouse 3's final maze completion time.
- D. More practice reduced each mouse's maze completion time by more than 10 seconds.

Item Information				Option Annotations
Alignment		S8.A.2.1.1		A. Key: Maze completion times decreased for all three mice with each trial. B. Practice did have an effect, as maze completion times decreased with each trial. C. Mouse 3 did not have the greatest decrease in maze completion times. D. Maze completion times for Mouse 2 did not decrease by more than 10 seconds.
Answer Key		A		
Depth of Knowledge		2		
p-values				
A	B	C	D	
63%	12%	13%	12%	

Use the graph below to answer question 5.



5. Which statement **best** describes the relationship between the length of a metal spring and its potential energy as shown in the graph?
- A. Increased potential energy in the spring forces it to extend.
 - B. Increased potential energy in the spring forces it to contract.
 - C. As the length of the spring increases, its potential energy increases.
 - D. As the length of the spring increases, its potential energy decreases.

Item Information				Option Annotations
Alignment		S8.A.2.1.4		A. Potential energy does not force the spring to extend. B. Potential energy does not force the spring to contract. C. Key: Increased spring length resulted in increased potential energy. D. Potential energy does not decrease as the spring length increases.
Answer Key		C		
Depth of Knowledge		2		
p-values				
A	B	C	D	
18%	5%	72%	5%	

6. Which measurements are required to calculate the average speed of a car driving on a road?
- A. time and distance
 - B. time and direction
 - C. mass and distance
 - D. mass and direction

Item Information				Option Annotations
Alignment		S8.A.2.2.2		A. Key: Time and distance are required to calculate speed. B. Time is used to calculate speed, but direction is not. C. Distance is used to calculate speed, but mass is not. D. Neither mass nor direction is used to calculate speed.
Answer Key		A		
Depth of Knowledge		2		
p-values				
A	B	C	D	
84%	5%	8%	3%	

7. Which part of a farm system is directly improved by the use of fossil-fuel-based fertilizers?
- A. air quality
 - B. water purity
 - C. soil temperature
 - D. nutrients and minerals

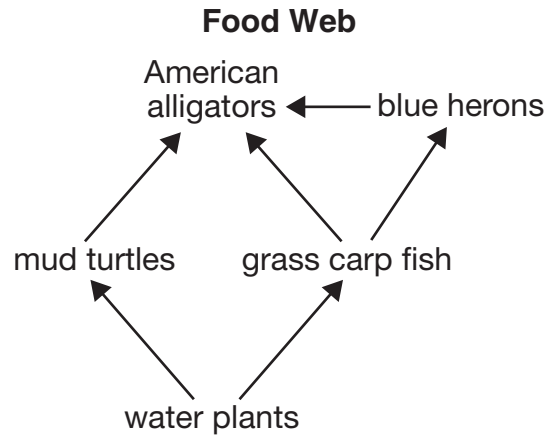
Item Information				Option Annotations
Alignment		S8.A.3.1.5		A. Air quality can be reduced by the use of fossil-fuel-based fertilizers. B. Water quality can be reduced by the use of fossil-fuel-based fertilizers. C. Soil temperature is not improved by the use of fossil-fuel-based fertilizers. D. Key: Fossil-fuel-based fertilizers provide nutrients and minerals to soils.
Answer Key		D		
Depth of Knowledge		2		
p-values				
A	B	C	D	
8%	8%	17%	68%	

8. Which word represents the basic unit of heredity in living organisms?

- A. gene
- B. zygote
- C. nucleus
- D. chromosome

Item Information				Option Annotations
Alignment		S8.B.2.2.2		A. Key: A gene is the basic unit of heredity. B. A zygote is formed by fertilization of an egg by a sperm. C. A nucleus is the control center of a cell. D. A chromosome is a strand of DNA that contains many genes.
Answer Key		A		
Depth of Knowledge		1		
<i>p</i> -values				
A	B	C	D	
62%	7%	12%	19%	

Use the food web below to answer question 9.



9. Which statement **best** describes one way energy flows through this food web?
- A. Energy flows from American alligators to mud turtles to water plants.
 - B. Energy flows from American alligators to blue herons to grass carp fish.
 - C. Energy flows from water plants to mud turtles to American alligators to grass carp fish.
 - D. Energy flows from water plants to grass carp fish to blue herons to American alligators.

Item Information				Option Annotations
Alignment		S8.B.3.1.1		A. Energy does not flow from consumers to producers. B. Energy does not flow from higher-level consumers to lower-level consumers. C. Energy does not flow from alligators to grass carp fish. D. Key: Energy from the Sun enters the food web through water plants and is then transferred to grass carp fish, to blue herons, and then to American alligators.
Answer Key		D		
Depth of Knowledge		2		
p-values				
A	B	C	D	
5%	4%	7%	84%	

Use the list below to answer question 10.

Organism Traits

1. Good night vision
2. Good depth perception
3. Very sensitive hearing
4. Fast, silent flight
5. Sharp talons
6. Hooked beaks

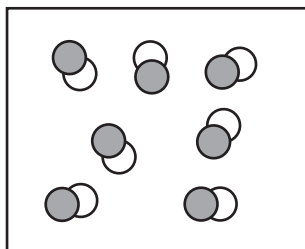
10. Which word **best** describes the organism's role in its ecosystem?

- A. mutualist
- B. parasite
- C. predator
- D. producer

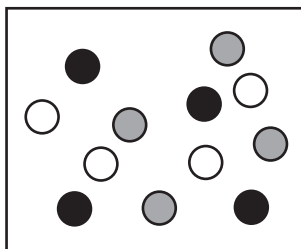
Item Information				Option Annotations
Alignment		S8.B.3.1.3		A. The traits in the list do not indicate living with another organism and both benefiting. B. The traits in the list do not indicate living with another organism at the expense of the other organism. C. Key: The traits in the list indicate characteristics that benefit hunting of other organisms. D. The traits in the list do not indicate absorbing sunlight to produce food.
Answer Key		C		
Depth of Knowledge		2		
p-values				
A	B	C	D	
5%	5%	84%	6%	

Use the diagrams below to answer question 11.

Model X



Model Y



11. Which statement **best** describes the model that shows a compound?

- A. Model X shows a compound because there are two different types of atoms.
- B. Model X shows a compound because two different atoms are chemically bonded together.
- C. Model Y shows a compound because there is more than one type of atom.
- D. Model Y shows a compound because each pair of matching atoms can form a chemical bond.

Item Information				Option Annotations
Alignment		S8.C.1.1.1		
Answer Key		B		
Depth of Knowledge		2		
p-values				
A	B	C	D	
8%	72%	9%	11%	

A. Model X shows a compound, but a compound is not defined by the presence of two different types of atoms.

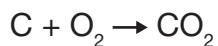
B. Key: Model X shows a compound, and compounds are atoms chemically bonded together.

C. Model Y does not show a compound, and a compound is not defined by the presence of more than one type of atom.

D. Model Y does not show a compound, and a compound is not defined by matching atoms forming a chemical bond.

Use the chemical equation below to answer question 12.

Formation of Carbon Dioxide



12. Which statement **best** describes the chemicals in the reaction?

- A. C and O₂ are reactants, and CO₂ is a product.
- B. C and O₂ are products, and CO₂ is a reactant.
- C. C and CO₂ are products because they contain carbon.
- D. C, O₂, and CO₂ are all reactants because they are involved in a reaction.

Item Information				<div>Option Annotations</div> <div>A. Key: In a chemical equation, the reactants are located to the left of the arrow, and the products are located to the right.</div> <div>B. C and O₂ are reactants, and CO₂ is a product.</div> <div>C. C is not a product.</div> <div>D. CO₂ is not a reactant.</div>
Alignment		S8.C.1.1.3		
Answer Key		A		
Depth of Knowledge		2		
p-values				
A	B	C	D	
64%	23%	6%	7%	

13. Which example is a nonrenewable source of chemical energy?

- A. biofuel
- B. windmill
- C. gasoline
- D. solar panel

Item Information				Option Annotations
Alignment		S8.C.2.1.1		A. Biofuel is a source of chemical energy, but it is renewable. B. Wind is not a source of chemical energy and is renewable. C. Key: Gasoline is a nonrenewable source of chemical energy. D. Sunlight is not a source of chemical energy and is renewable.
Answer Key		C		
Depth of Knowledge		2		
<i>p-values</i>				
A	B	C	D	
18%	6%	70%	6%	

14. Which statement **best** explains the importance of fossils to scientists?

- A. Fossils show how animals viewed their surroundings, so scientists know more about past animals.
- B. Fossils show what color animals once were, so scientists know more about what they ate.
- C. Fossils show where animals once lived, so scientists know more about the environment and how it has changed.
- D. Fossils show that animals lived in the same location today as they once did, so scientists know more about today's environment.

Item Information				Option Annotations
Alignment		S8.D.1.1.4		A. Fossils do not involve the viewpoint of animals. B. An animal’s color is seldom preserved and does not provide information on food sources. C. Key: The type of organism that formed the fossil can provide insight to past environments based on the type of resources the organism needed to survive. D. Fossils do not tell us about current environments.
Answer Key		C		
Depth of Knowledge		2		
p-values				
A	B	C	D	
14%	4%	74%	8%	

15. The cycling of water from ocean surfaces to the atmosphere mainly depends on which process?
- A. evaporation caused by the Sun
 - B. precipitation caused by the Sun
 - C. evaporation caused by warm ocean currents
 - D. precipitation caused by warm ocean currents

Item Information				Option Annotations
Alignment		S8.D.1.3.1		A. Key: Energy from the Sun increases the movement of water molecules which causes evaporation of water, which rises from ocean surfaces to the atmosphere. B. Precipitation moves water from the atmosphere to the ocean. C. Warm ocean currents are not the main factor that drives evaporation from the ocean’s surface to the atmosphere. D. Precipitation moves water from the atmosphere to the ocean.
Answer Key		A		
Depth of Knowledge		1		
p-values				
A	B	C	D	
76%	8%	9%	6%	

Use the drawing below to answer question 16.

951 Gaspra



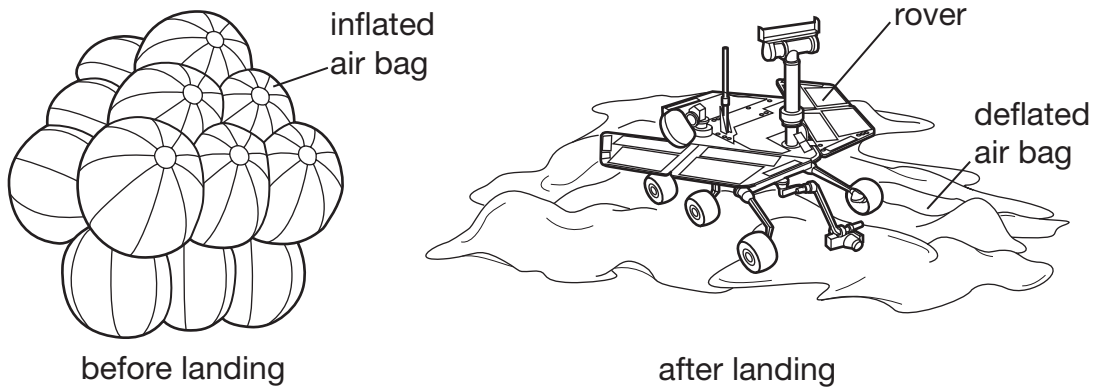
16. The drawing shows the asteroid 951 Gaspra as seen by the Galileo spacecraft as it passed through the asteroid belt. How is this asteroid different from a moon?
- A. Unlike a moon, Gaspra has a gravitational force.
 - B. Unlike a moon, Gaspra is made of solid materials.
 - C. Gaspra would be classified as a moon if it were close to a planet.
 - D. Gaspra would be classified as a moon if it began orbiting a planet.

Item Information				Option Annotations
Alignment		S8.D.3.1.3		A. All objects with mass have gravity. B. Moons are also made of solids. C. Proximity to a planet is not what classifies a body as a moon. D. Key: A moon remains in orbit around a planet.
Answer Key		D		
Depth of Knowledge		2		
p-values				
A	B	C	D	
18%	13%	6%	63%	

OPEN-ENDED QUESTIONS

Use the diagram below to answer question 17.

Mars Exploration Rover Landing System



17. The landing system for the Mars Exploration Rover spacecraft used a system of air bags to protect its fragile rover during landing. Engineers first used a computer simulation to test the new air bag design; then they used crash tests with various types of vehicles.

Part A: Describe one benefit of using a computer simulation to test the air bag design before conducting crash tests.

Part B: Explain how crash tests help engineers test the air bag design.

SCORING GUIDE

#17 ITEM INFORMATION

Alignment	S8.A.3.2.2	Depth of Knowledge	2	Mean Score	1.35
------------------	------------	---------------------------	---	-------------------	------

ITEM-SPECIFIC SCORING GUIDELINE

Score	Description
2	<p>The response demonstrates a <i>thorough</i> understanding of how engineers use models to develop new and improved technologies to solve problems by</p> <ul style="list-style-type: none"> describing one benefit of using a computer simulation to test the air bag design before conducting crash tests <p>AND</p> <ul style="list-style-type: none"> explaining how crash tests help engineers test the air bag design. <p>The response is clear, complete, and correct.</p>
1	<p>The response demonstrates a <i>partial</i> understanding of how engineers use models to develop new and improved technologies to solve problems by</p> <ul style="list-style-type: none"> describing one benefit of using a computer simulation to test the air bag design before conducting crash tests <p>OR</p> <ul style="list-style-type: none"> explaining how crash tests help engineers test the air bag design. <p>The response may contain some work that is incomplete or unclear.</p>
0	<p>The response provides <i>insufficient</i> evidence to demonstrate any understanding of the concept being tested.</p>
Non-scorables	<p>B – No response written R – Refusal to respond F – Foreign language K – Off task U – Unreadable</p>

Note: No deductions should be taken for misspelled words or grammatical errors.

Responses that will receive credit:**Part A (1 point):**

- One benefit of using a computer simulation to test the air bag design before a crash test is safety.
- Computer simulations can save money and/or resources during product development.
- Computer simulations benefit the testing process by modeling conditions in which equipment will operate that are not present on Earth (i.e., conditions on Mars).
- Computer simulations make the testing process easier by allowing for different calculations to be made and tested quickly and efficiently.
- Computer simulations make it possible to collect large amounts of data in a short time and at a lower cost than conducting crash tests.

Part B (1 point):

- Crash tests help engineers test the air bag design by allowing for direct observation.
- Crash tests allow engineers to put theory into practice by observing how materials within a product perform.
- Crash tests allow engineers to study landing gear materials after a crash to identify potential weaknesses in the materials, their production, or how the landing process was executed.

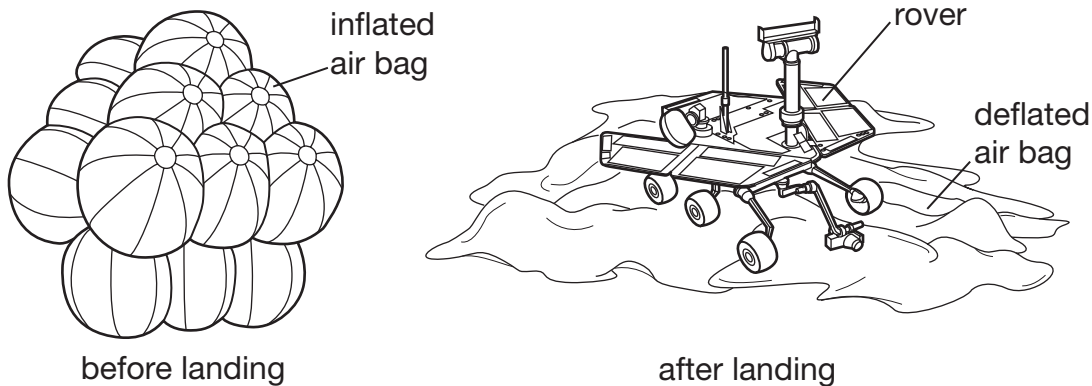
**THIS PAGE IS
INTENTIONALLY BLANK.**

STUDENT RESPONSE

RESPONSE SCORE: 2 POINTS

Use the diagram below to answer question 17.

Mars Exploration Rover Landing System



17. The landing system for the Mars Exploration Rover spacecraft used a system of air bags to protect its fragile rover during landing. Engineers first used a computer simulation to test the new air bag design; then they used crash tests with various types of vehicles.

Part A: Describe one benefit of using a computer simulation to test the air bag design before conducting crash tests.

A benefit is that if you know it isn't going to work on the computer, you don't waste materials.

Part B: Explain how crash tests help engineers test the air bag design.

It lets them know if the air bag is strong enough and know if it's going to pop or not.

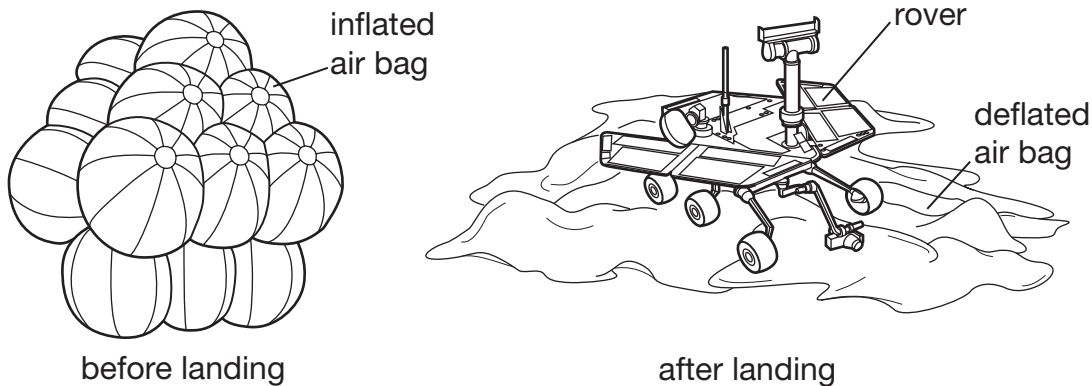
This response demonstrates a thorough understanding of how engineers use models to develop new technologies to solve problems by describing a benefit of computer simulations ("if you know it isn't going to work on the computer you don't waste materials") and by explaining how crash tests help test the airbag design ("know if it's going to pop or not"). The response is clear, complete, and correct.

STUDENT RESPONSE

RESPONSE SCORE: 1 POINT

Use the diagram below to answer question 17.

Mars Exploration Rover Landing System



17. The landing system for the Mars Exploration Rover spacecraft used a system of air bags to protect its fragile rover during landing. Engineers first used a computer simulation to test the new air bag design; then they used crash tests with various types of vehicles.

Part A: Describe one benefit of using a computer simulation to test the air bag design before conducting crash tests.

You will have an idea of what will happen.

Part B: Explain how crash tests help engineers test the air bag design.

If it doesn't work good then they would change it.

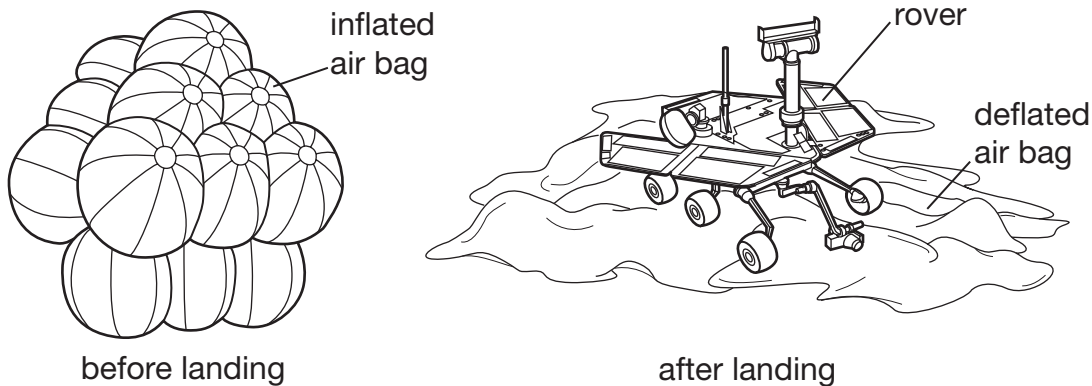
This response demonstrates a partial understanding of how and why engineers use models to develop new technologies to solve problems. The response to Part A is too vague for credit because it fails to describe a specific benefit of running a computer simulation. The response to Part B, "If it doesn't work good then they would change it," is minimal, but acceptable, indicating the design can be improved based on crash test results.

STUDENT RESPONSE

RESPONSE SCORE: 0 POINTS

Use the diagram below to answer question 17.

Mars Exploration Rover Landing System



17. The landing system for the Mars Exploration Rover spacecraft used a system of air bags to protect its fragile rover during landing. Engineers first used a computer simulation to test the new air bag design; then they used crash tests with various types of vehicles.

Part A: Describe one benefit of using a computer simulation to test the air bag design before conducting crash tests.

They will know if the product works or not.

Part B: Explain how crash tests help engineers test the air bag design.

They can use vehicles that arnt important to test things for vehicles that are important.







This response provides insufficient evidence to demonstrate understanding of the concepts being tested. "They will know if the product works or not" does not describe a benefit of using computer simulations prior to conducting crash tests. The response in Part B is unclear. The use of "important" vs. unimportant vehicles does not explain how crash tests help engineers test their design.

WBTE Preview

Question 18

Albert Einstein

?

Back

Describe two ways that animals respond to the seasonal weather changes in Pennsylvania.

1. Eq

0 / 1000

2. Eq

0 / 1000

Review/End Test

Pause

Flag

Options

SCORING GUIDE

#18 ITEM INFORMATION

Alignment	S8.B.3.2.3	Depth of Knowledge	2	Mean Score	1.50
-----------	------------	--------------------	---	------------	------

ITEM-SPECIFIC SCORING GUIDELINE

Score	Description
2	<p>The response demonstrates a <i>thorough</i> understanding of the response of organisms to environmental changes (e.g., changes in climate, hibernation, migration, coloration) by describing two ways that animals respond to the seasonal weather changes in Pennsylvania.</p> <p>The response is clear, complete, and correct.</p>
1	<p>The response demonstrates a <i>partial</i> understanding of the response of organisms to environmental changes (e.g., changes in climate, hibernation, migration, coloration) by describing one way that animals respond to the seasonal weather changes in Pennsylvania.</p> <p>The response may contain some work that is incomplete or unclear.</p>
0	The response provides <i>insufficient</i> evidence to demonstrate any understanding of the concept being tested.
Non-scorables	<p>B – No response written</p> <p>R – Refusal to respond</p> <p>F – Foreign language</p> <p>K – Off task</p> <p>U – Unreadable</p>

Note: No deductions should be taken for misspelled words or grammatical errors.

Responses that will receive credit:

Ways that animals respond (1 point each):

- Some animals hibernate, go into torpor, or become dormant in the winter.
- Some animals migrate to a warmer climate in the fall and return in the spring.
- Some animals' fur turns a lighter color (white) in winter and turns darker in spring.
- Some animals grow thicker fur / fat layers in preparation for winter.
- Some animals store food in the fall.
- Some animals find winter shelter.

STUDENT RESPONSE

RESPONSE SCORE: 2 POINTS

WBTE Preview Albert Einstein

Question 18

Describe two ways that animals respond to the seasonal weather changes in Pennsylvania.

1. Birds fly south for the winter so they can be in warmer weather and not cold here.

2. Mice travel into houses and people places because it is so cold during the winter.

Review/End Test Pause Flag Options Back

This response demonstrates a thorough understanding of the response of organisms to environmental changes by describing two different responses. "Birds fly south" and "mice travel into houses and people places" are acceptable descriptions of animal responses to a seasonal weather change. The response is clear, complete, and correct.

STUDENT RESPONSE

RESPONSE SCORE: 1 POINT

WBTE Preview **Albert Einstein**

Question 18

Describe two ways that animals respond to the seasonal weather changes in Pennsylvania.

1. Some mammals hibernate. That means they fall asleep during winter.

2. Other animals adapt to the different seasons every season.

This response demonstrates a partial understanding of the response of organisms to environmental changes. "Some mammals hibernate" is an acceptable description of an animal response. "Other animals adapt to the different seasons" is too vague for credit. This response contains some work that is incomplete and unclear.

STUDENT RESPONSE

RESPONSE SCORE: 0 POINTS

WBTE Preview Albert Einstein

Question 18

Describe two ways that animals respond to the seasonal weather changes in Pennsylvania.

1. the animals know what to do during the kind of season.

2. they respond differently then humans do due to the weather.

Review/End Test Pause Flag Options Back

This response provides insufficient evidence to demonstrate any understanding of the response of organisms to environmental changes. "The animals know what to do" does not describe specifically what it is that they do. The student fails to provide descriptions of two ways animals respond to seasonal changes, and the responses given do not demonstrate any understanding of the concept being tested.

SCIENCE GRADE 8—SUMMARY DATA

MULTIPLE-CHOICE

Sample Number	Alignment	Answer Key	Depth of Knowledge	<i>p</i> -values			
				A	B	C	D
1	S8.A.1.1.3	D	2	10%	4%	16%	70%
2	S8.A.1.2.4	D	2	14%	13%	7%	66%
3	S8.A.1.3.1	C	2	10%	9%	74%	7%
4	S8.A.2.1.1	A	2	63%	12%	13%	12%
5	S8.A.2.1.4	C	2	18%	5%	72%	5%
6	S8.A.2.2.2	A	2	84%	5%	8%	3%
7	S8.A.3.1.5	D	2	8%	8%	17%	68%
8	S8.B.2.2.2	A	1	62%	7%	12%	19%
9	S8.B.3.1.1	D	2	5%	4%	7%	84%
10	S8.B.3.1.3	C	2	5%	5%	84%	6%
11	S8.C.1.1.1	B	2	8%	72%	9%	11%
12	S8.C.1.1.3	A	2	64%	23%	6%	7%
13	S8.C.2.1.1	C	2	18%	6%	70%	6%
14	S8.D.1.1.4	C	2	14%	4%	74%	8%
15	S8.D.1.3.1	A	1	76%	8%	9%	6%
16	S8.D.3.1.3	D	2	18%	13%	6%	63%

OPEN-ENDED

Sample Number	Alignment	Points	Depth of Knowledge	Mean Score
17	S8.A.3.2.2	2	2	1.35
18	S8.B.3.2.3	2	2	1.50

PSSA Grade 8 Science Item and Scoring Sampler

Copyright © 2015 by the Pennsylvania Department of Education. The materials contained in this publication may be duplicated by Pennsylvania educators for local classroom use. This permission does not extend to the duplication of materials for commercial use.



Item Writer Manual

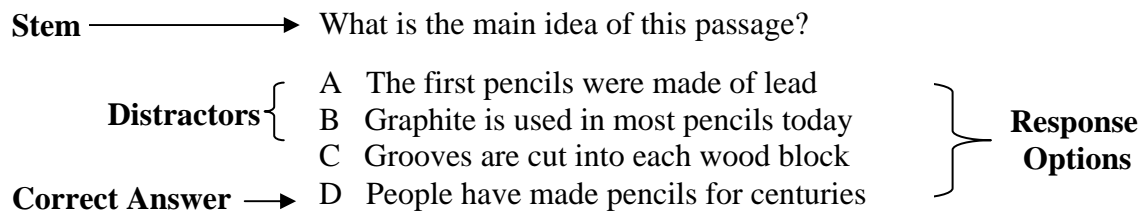
Table of Contents

Models of Multiple Choice and Constructed-Response Items.....	3
Structure of Multiple Choice Stems.....	5
Structure of Multiple Choice Response Options.....	11
Content of Multiple Choice Items.....	18
Arrangement of Items in Test.....	21
Pitfalls to Avoid.....	23
Style Issues.....	29
Bias and Sensitivity.....	30
Topics to Avoid.....	31
Exceptions.....	33
Glossary of Assessment Terms.....	34

Models of Multiple Choice and Constructed-Response Items

State assessments commonly use two types of items: multiple choice (MC) and constructed response (CR).

Multiple Choice Model



Definitions of Terms

Stem: the question or incomplete statement that establishes a problem

Response Options: answer choices

Correct Answer (CA): key

Distractors: incorrect answers or foils

Constructed-Response Model

Prompt → Think of two words that describe the first little pig, and explain why they are good choices.

Response → The first little pig is lazy and foolish. He is lazy because he builds his house out of straw so he can finish quickly. He is foolish because a straw house is easy to destroy.

Rubric	Score 4	The response includes two appropriate words, and each is explained logically.
	Score 3	The response includes two appropriate words but only one is explained logically.
	Score 2	The response includes two appropriate words without logical explanations, or it includes one appropriate word explained logically.
	Score 1	The response includes one appropriate word without logical explanation.

Definitions of Terms

Prompt: the question or direction that poses a problem for students to answer or solve

Response: the student's answer

Rubric: the guidelines for scoring student responses

Structure of Multiple Choice Stems

- 1 Stems may be open or closed.** States will indicate the type of stem(s) they prefer. An open stem is an incomplete sentence that is completed by the appropriate response option. A closed stem is simply a direct question, ending with a question mark.

Examples (Correct responses are indicated by an asterisk.)

(Students have read a story about a giant fish in which the fisher throws the giant fish back into the ocean because it has caused the fisher problems.)

A. Open Stem

The fisher throws the giant fish back into the ocean because —

- A the giant fish has brought the fisher nothing but trouble*
- B the fisher wants someone else to catch the giant fish in the future
- C the giant fish wants to return to the ocean
- D the fisher wants to try and catch the giant fish again

B. Closed Stem

Why does the fisher throw the giant fish back into the ocean?

- A The giant fish has brought the fisher nothing but trouble.*
- B The fisher wants someone else to catch the giant fish in the future.
- C The giant fish wants to return to the ocean.
- D The fisher wants to try and catch the giant fish again.

These items are very similar. However, the closed stem is somewhat better because it states the problem explicitly. Also, it is easier for students to remember a complete question, so less capable students will not have to struggle to hold the stem in mind as they test each possible choice against it.

Although closed stems are usually preferred, open stems may work better in some situations.

Examples

(Students have read an article about Groundhog Day.)

A. Closed Stem

During which month does the groundhog first poke its head out of its burrow?

- A January
- B February*
- C March
- D April

B. Open Stem

The groundhog first pokes its head out of its burrow in —

- A January
- B February*
- C March
- D April

Here the open stem is better because it is more streamlined and thus easier to read.

- 2 Each stem, whether it is open or closed in structure, should present one clearly stated problem. Students should be able to identify the problem from the stem alone. If they cannot do so, the item is flawed.

Examples

(Students have read a passage about a girl who has a friendly dog.)

A. Poor

Rosa--

- A plays baseball
- B is in eighth grade
- C loses her notebook
- D has a friendly dog*

B. Better

Rosa's dog is **best** described as—

- A shy
- B lazy
- C clever
- D friendly*

The first stem is inadequate because no problem is presented and no question is asked. Example B clearly prompts students to select the word that best describes the dog.

- 3 Include information in the stem to avoid repeating it in each option.** This is an important technique for eliminating unnecessary words, but it is not an ironclad rule. Here and elsewhere, good judgment is required.

Examples

- A. Poor
The Monroe Doctrine established the U.S. policy against —
- A new European colonies in Southeast Asia
B new European colonies in Latin America*
C new European colonies in North Africa
D new European colonies in South Africa
- B. Better
The Monroe Doctrine established the U.S. policy against new European colonies in —
- A Southeast Asia
B Latin America*
C North Africa
D South Africa

There is no reason for students to read the words “new European colonies in” four times. The test should assess the students’ ability to comprehend the items—not to plow through redundancy in the item choices. Example B is more clearly stated. This is an important rule for writing concise items, but there are exceptions.

Examples

- A. Poor
One reason that whales are classified as mammals rather than as fish is because whales —
- A breathe in air
B have bony skeletons
C produce milk to feed their young*
D cannot live out of the water
- What is one reason that whales are classified as mammals rather than fish?
- A They breathe in air.
B They have bony skeletons.
C They produce milk to feed their young.*
D They cannot live out of the water.

B. Better

Example A exhibits a problem that sometimes occurs with an open stem. Students who are not strategic readers may try to hold the stem in memory while they test it against each option. In Example B, the closed stem clearly states a memorable question. Though repeated four times, word “They” does not cause a significant increase in the reading load. The four response options are clear and complete sentences.

- 4 **Stems, like all parts of a test item, should be clear and concise.** No matter what content area the item covers, it should test content knowledge and/or content area processing and thinking skills—not test-taking skills.
- 5 **Use the active voice.** Items written in the active voice usually sound the most fluent and interesting. Try to avoid passive voice. Present tense is preferred; use past tense only for historical items.
- 6 **Use grade-level-appropriate vocabulary.** The reading difficulty should come from understanding the passages, not from understanding the item stems. For some state projects, a list or booklet of grade-level-appropriate words may be provided. Such lists can be helpful guides. However, when they are followed too slavishly, they can be frustrating impediments to good test development.

Examples

(Students have read the a passage about the White House and its residents.)

A. Poor

The first chief executive to preside
for two consecutive terms was —

- A George Washington*
- B John Adams
- C Thomas Jefferson
- D James Madison

B. Better

The first president to serve for two
terms in a row was —

- A George Washington*
- B John Adams
- C Thomas Jefferson
- D James Madison

Example B uses grade level appropriate vocabulary and tests the student's comprehension of the passage.

- 7 Stems and options should include only what is necessary.** Resist the temptation to ask two or more questions in one item or to teach while testing.

Examples

(Students have read a passage about a boy named Javan who does kind things for other children, who do kind things for Javan in return.)

A. Poor

What kind act does Javan perform on Tuesday, and how does it benefit him?

- A He shares his lunch with Martha, and she gives Javan a drawing.
- B He waits for Paul, and Paul helps Javan clean his room.
- C He saves a seat for Anne, and she invites Javan to a picnic.*
- D He carries Jamal's books, and Jamal lets Javan borrow one.

B. Better

Anne invites Javan to a picnic to thank him for —

- A sharing his lunch
- B waiting for her
- C saving her a seat*
- D carrying her books

Example B clearly states one problem and is more focused. The events used in the poor distractors can now be used in other items without cueing/clueing or overlapping with this item.

- 8 Each item should measure only one standard.** This point applies to all content areas, but it can often be seen more easily in certain subtests such as Language Mechanics.

Examples

Select the sentence that is written correctly.

A. Poor

- A After the reign stopped and the son came out.
- B Latasha bought apples grapes, pears, and bananas.
- C The class picnic will be held in Russell Park.*
- D Ming will be late, he has to clean his room.

B. Better

- A The class picnic on the last day of school.
- B A chance to celebrate the end of the school year.
- C This has been a tradition for many years.*
- D With food, games, and prizes for everyone.

In Example A, each response option addresses a different standard. Choice A tests the correct spelling of two homophones in context and recognition of sentence fragments; Choice B tests correct use of commas in a series; Choice C contains no errors; and Choice D tests recognition of a run-on sentence.

In Example B, all of the distractors are sentence fragments, and the keyed answer is a correctly written sentence. If a student answers Example B correctly, it is reasonable to conclude that the student demonstrated the ability to recognize and avoid sentence fragments. If a student answers Example A correctly, it is difficult to determine what the student has demonstrated.

Structure of Multiple Choice Response Options

- 1 **Include *only one* correct answer for each stem.** If a distractor is defensible, the item is not fair. The right answer should be clearly correct, and each of the wrong answers should be indefensible.

Examples

(Students have just read a passage about European exploration of the New World in which several motives for such exploration have been given.)

A. Poor

Which was the **most** important reason for Spanish exploration of the New World?

- A finding gold
- B winning glory
- C spreading religion
- D enlarging the empire

B. Better

Which resource of the New World was **most** important to Spanish explorers in the 1500s?

- A gold*
- B soil
- C lumber
- D furs

Example A is poor because all of the response options are defensible. Unless students are explicitly told in the passage that one reason is more important, all answers are defensible, so there is no correct answer. If one reason *has* been described in a passage as most important, the item stem should be edited to read, “According to this passage, which was the **most** important reason for Spanish exploration of the New World?” Example B is acceptable as is for a social studies test if it matches a curriculum standard.

More Examples

(On a mathematics test, students are asked to identify shapes.)

A. Poor

Figure X is a —

- A triangle
- B square*
- C rectangle
- D pentagon

A. Better

Figure X is a —

- A triangle
- B square*
- C pentagon
- D hexagon

Because *all* squares are also rectangles, the Example A has two correct answers. Even adding the highlighted word **best** to the stem will not make these response options acceptable.

- 2 Within an item, all of the response options should be parallel in structure and content.** This means that all options should be generally the same length, the same level of abstraction, and (in most cases in which verbal options are used) contain the same part of speech or the same grammatical structure.

Examples

(Students have read a passage about a boy who is bored because he cannot find anyone to play with him.)

A. Poor

Why does Mike feel bored at the beginning of the story?

- A He has no children in his neighborhood to play with.*
- B It is raining.
- C Nothing is on television.
- D He is hyper.

In Example A, the correct answer is a long and plausible sentence, while option B is short, option C is a sentence fragment, and option D is short, negative, and inappropriate. If test-wise students can select option A without reading the passage and, in this case, without even reading the question, the item is seriously flawed.

B. Better

Why does Mike feel bored at the beginning of the story?

- A He has no one to play with.*
- B He has to stay in his room.
- C Everyone else in his family is busy.
- D Rain has kept him indoors all week.

In Example B, all of the choices are plausible, of similar length, and are appropriate possibilities. They do not need to be exactly the same. The rules for acceptable parallelism vary from client to client. Some state clients consider the options parallel only if all are the same, while other clients require that they all be different. Still other state clients might require that two options be the same in one way, and two the same in another way. In this example, two are slightly shorter, and two are slightly longer. Some clients are more flexible than others. The most important point is that the correct answer should not stand out from the other three options.

- 3 No response option should contradict or negate information presented in the stem.** This would be an easily spotted throwaway option.

Examples

(Students have read a story about a family that has moved to a new apartment building.)

A. Poor

Why did Hiroshi's family move to Maplewood Towers?

- A They did not move because no pets are allowed in Maplewood Towers.
- B They wanted to live closer to Hiroshi's new school.
- C They had many friends living in Maplewood Towers.
- D They liked the view from their new apartment.

B. Better

Why did Hiroshi's family move to Maplewood Towers?

- A They wanted to live in the same building as Ray's grandmother.
- B They wanted to live closer to Hiroshi's new school.
- C They had many friends living in Maplewood Towers.
- D They liked the view from their new apartment.

Students should be able to trust the information in the stem as correct. In Example A, students who understand that the information in the stem must be true now have only three viable options instead of four. Choice A becomes a throwaway option.

4 Response options should be as brief as possible. Lengthy answers are undesirable.

5 All of the options must fit grammatically and syntactically with the stem.

Students should not be able to select or eliminate any options because of grammar or syntax.

Examples

(Students have read a passage about a man who buys his daughter a snack.)

A. Poor

Mr. Jackson gives Carla a—

- A pretzel
- B hotdog
- C slice of pie
- D ice cream cone

B. Better

Mr. Jackson gives Carla—

- A a pretzel
- B a hotdog
- C an apple raisin roll
- D an ice cream cone

In Example A, test-wise students can eliminate choice D because it does not follow the stem grammatically.

6 Provide at least three plausible distractors. In some cases, there cannot logically be enough plausible distractors.

Examples

A. Poor

During times of inflation, the prices of goods generally —

- A rise*
- B fall
- C stay the same
- D rise and then fall

B. Better

What usually happens during a period of inflation?

- A Sales decrease.
- B Prices increase.*
- C Job openings decrease.
- D Personal savings increase.

In Example A, options A and B are the two most plausible choices. Option C is barely plausible, and option D is even less plausible. In Example B, the options all look plausible (to students who do not know the answer), and they are also parallel.

7 All of the response options should be drawn from the same text or from thematically related content.

Examples

(Students taking a social studies test are asked a question about the branches of government.)

A. Poor

Which is a basic role of Congress?

- A making movies
- B commanding the military
- C passing new laws*
- D making rulings at trials

B. Better

Which is a basic role of Congress?

- A enforcing laws
- B commanding the military
- C passing new laws*
- D making rulings at trials

In Example A, test-wise students can eliminate option A because it is not a role of government. Therefore, they have a better than one-in-four chance of guessing the correct response, even if they do not know the roles of the branches of government.

More Examples

(Students have read a story about a child's day at school.)

A. Poor

What does Antonio leave at school by mistake?

- A a coat
- B a book
- C his report card*
- D his dog's bowl

B. Better

What does Antonio leave at school by mistake?

- A a coat
- B a book
- C his report card*
- D his backpack

In Example A, test-wise students can eliminate option D because it is not something generally associated with going to school. Ideally, all of the options for this kind of item should be mentioned in the passage. If the dog's bowl is mentioned in the story, that choice becomes slightly more plausible. If Antonio has brought the dog's bowl to school for some reason, then the option is acceptable.

In other words, do not bring in response options from left field. Use options that appear in a passage or stimulus or ones that make sense in the established context.

- 8 If options are numbers, times, dates, or other quantitative or sequential ideas, they should generally be arranged in either ascending or descending order (usually ascending).**

Examples

(Students taking a social studies test are presented with a bar graph stimulus.)

Rainfall amounts for this region were greatest in —

A. Unacceptable Order:

- A 2003
- B 2000
- C 2002
- D 2001

B. Acceptable Order:

- A 2000
- B 2001
- C 2002
- D 2003

Once students have used the bar graph to determine the answer, they should not have to hunt among the options to find it. Although locating the answer may seem to be a simple task, hunting among the options may cause some students who have identified the correct response to mark a different option. This rule generally applies to any sequential response options, including days of the week, months of the year (especially when they are consecutive), amounts of money, lengths, weights, etc. The majority of items on a mathematics test will be governed by this rule, with some exceptions.

Exception

This item might appear on a mathematics test.

A. Unacceptable

Which has the greatest value?

- A $\frac{1}{2}$
- B $\frac{2}{3}$
- C $\frac{3}{4}$
- D $\frac{4}{5}^*$

B. Acceptable

Which has the greatest value?

- A $\frac{3}{4}$
- B $\frac{2}{3}$
- C $\frac{4}{5}^*$
- D $\frac{1}{2}$

In this case, arranging the options in ascending order would give away the answer.

- 9 Avoid using absolute words.** Words such as “all,” “none,” “always,” and “never” are red flags to test-wise students. Using these absolutes often results in options that students can easily eliminate.

Examples

A. Poor

What happens during a period of inflation?

- A People always buy less.
- B Prices tend to increase.*
- C No jobs are available.
- D People save all their money.

B. Better

What usually happens during a period of inflation?

- A Most people buy less.
- B Prices tend to increase.*
- C Few jobs are available.
- D People save more of their money.

In Example A, options A, C, and D can be eliminated because students know that there are exceptions.

Content of Multiple Choice Items

- 1 Compose items that allow students to show their understanding of the curriculum being tested.** State standards will indicate the types of knowledge, concepts, and skills that are being measured. Link each item to a standard.
- 2 Test knowledge and skills that are important.** Avoid testing knowledge of trivial facts. Questions that require higher-level thinking are generally desirable. (See *Bloom's Taxonomy* or other sources for more information about higher-level thinking.) If a standard calls for locating or recalling specific facts or details, items should address *important* facts or details.

Examples

(Students have read a passage about the White House and its most famous First Ladies.)

A. Poor

In which year did British troops set fire to the White House?

- A 1792
- B 1814*
- C 1902
- D 1945

B. Better

Which First Lady saved White House paintings from a fire?

- A Martha Washington
- B Dolley Madison*
- C Eleanor Roosevelt
- D Jacqueline Kennedy

Most people would consider Dolley Madison's courageous act to be a more important fact than recalling the year in which it occurred. (Note that all of the response options are significant dates in the history of the White House that would have appeared in the passage.)

- 3 Include items for all standards being measured.** Compose items for an array of standards instead of concentrating on a favored few.

- 4 Provide passage-dependent items for reading tests.** An item is **passage-independent** if students can answer it without reading the passage. The assumption is that the information in the article or the plot of the story is unfamiliar to most students. Test developers need not be concerned about the exceptional student who is an expert on many subjects or about the voracious readers who may have already read a previously published passage. However, avoid items that significant numbers of students would be able to answer from prior knowledge or from a quick glance at the title.

Examples

(Students have read a passage called “George Washington: Boyhood to Manhood.”)

A. Poor

This passage is mainly about —

- A Martha Washington
- B George Washington*
- C the thirteen colonies
- D the American Revolution

B. Better

This passage is mainly about George Washington’s —

- A education
- B character*
- C appearance
- D wealth

Though both examples test the main idea standard, the correct answer to Example A is obvious from the title of the passage. Example B is more likely to require comprehension of the passage.

More Examples

A. Poor

Who was the first President of the United States?

- A John Adams
- B George Washington*
- C Thomas Jefferson
- D James Madison

B. Better

Who was George Washington’s vice-president?

- A James Monroe
- B John Adams*
- C James Madison
- D Thomas Jefferson

Most students could answer Example A from prior knowledge. At lower grades, few students could answer the second question without having read the passage. (Note: This is a somewhat lower-level, locating-information item. To ensure that the choices are all plausible, all of the individuals named in the response options should be mentioned in the passage.)

Science Examples

(Students observe a food web in which the sun provides the energy, which is transferred in turn to the wheat, to the mouse, to the snake, and to the hawk.)

A. Poor

The producer in this food web is the—

- A sun
- B wheat*
- C mouse
- D hawk

B. Better

[Same stem but a marine web]

- A sun
- B plankton*
- C minnow
- D frog

Example A is overused. Example B is a higher level thinking question that requires students to use information learned and to apply it to a similar situation.

5 Graphs, charts, maps, and other artwork require special attention. Graphic stimuli can often be used to support higher-order thinking and process skills, as well as make the test booklet pages look more interesting and engaging. Care and thought must be given when selecting graphic stimuli. Modifying the graphic may be necessary if something is missing. Here are some things to consider when selecting graphic stimuli:

- The question should be dependent upon the use of the graph, chart, or other artwork.
- All stimuli should be clear and simple for reproduction.
- All parts of the stimulus should be labeled properly.
- The content of the stimulus should be checked for accuracy and be current.
- Tables and other graphics often need titles; horizontal and vertical axes should be labeled using a consistent style; maps should have a compass rose etc.
- All information needed to answer the question should be provided in the stimulus.

An important distinction should be made between “decorative” and “functional” art. Decorative art gives little if any help to the test taker trying to understand a passage or an item. Functional art helps students understand and respond to a passage or an item.

Arrangement of Items in Test

- 1 Most tests should have both “floor” and “ceiling.”** In other words, there should be at least a few easy questions, and they should generally appear at or near the beginning of the test. There should also be at least a few difficult questions, and they should generally appear later in the test.

During field testing, item difficulties are not known. Editors can only approximate what they will be. After field testing, empirical data will indicate the exact difficulty of each item.

Many state clients now prefer to “spiral” the items; that is, to mix items with varying difficulty throughout the test. In the past, items were often arranged from least difficult to most difficult. On a reading test, these concerns are somewhat out of the test makers’ control because the items are grouped by passage.

- 2 Answer keys should be sensible but not predictable.** Each of the options should be used approximately—but not exactly—as much as the others. The correct answer can be in the same position two or three times in a row, but not much more than that. The answer key should not spell out any kind of message or constitute a pattern.

Examples

A. Poor 1

1 B
2 A
3 C
4 C
5 C
6 C
7 B
8 D
9 A
10 C

B. Poor 2

1 A
2 B
3 C
4 D
5 D
6 C
7 B
8 A
9 A
10 B

C. Poor 3

1 B
2 A
3 D
4 D
5 A
6 D
7 C
8 A
9 B
10 B

In Example A, the correct response occurs four times in a row. Twice in a row is fine, and three times is acceptable on occasion. Four or more is excessive. Also, note that D is correct only one time in 10.

In Example B, the letters are arranged sequentially from A to D and D to A. Students who know some of the answers can spot this type of pattern.

In Example C, each group of three letters, beginning with 1-3, forms a one-syllable word.

In general, answer keys should defy formulas. Each choice should be used *approximately* 25% of the time, but not exactly 25% of the time. You may have heard such axioms as “When in doubt, choose C” or “D is almost never correct.” Answer keys should **not** follow any axioms.

Pitfalls to Avoid

- 1 Do not assume a wide body of common knowledge.** Since common knowledge is never 100% common, getting an item right should not depend on having some background information that might not be accessible to a significant number of students.

Examples

(Students have read a story about a character that does not get a part in a play, so he pretends he never wanted one.)

A. Poor

This story is most like —

- A “The Princess and the Pea”
- B “The Boy Who Cried Wolf”
- C “The Fox and the Grapes”*
- D “The Mouse and the Lion”

B. Better

Why does Vijay say, “I never wanted to be in the play anyway”?

- A The play is boring to him.
- B He prefers other activities.
- C He is hiding his disappointment.*
- D Being in the play is hard work.

Example A assumes that students are familiar with four other stories outside of the passage. Although it requires higher-level thinking to make thematic comparisons, we are not interested in testing this outside knowledge. Example B is superior as it asks the students to deduce Vijay’s unstated motive for saying what he does.

- 2 Avoid idiomatic expressions that could be unfamiliar to students, especially students with an ESL background.** Although this is often treated as a bias issue, it is of general importance because students from all backgrounds should have an equal opportunity to answer questions correctly and should not be advantaged or disadvantaged by familiarity with idiomatic expressions.

Examples

(Students have read a story about a character named Ira who proves his loyalty.)

A. Poor

Beth thinks highly of Ivar because he is a —

- A quick study
- B cool customer
- C stand-up guy*
- D jack of all trades

B. Better

Beth thinks highly of Ivar because he—

- A learns new things quickly
- B stays calm under pressure
- C is loyal to his friends*
- D has many skills

Example A assumes that students are familiar with expressions that are either slang or idiomatic.

- 3 **Never use throwaway response options.** A throwaway option is so clearly wrong that most students will not even consider it as a possibility. An option may be implausible for several reasons.

Examples

A. Poor

Sarah's mother is **best** described as —

- A wise*
- B cruel
- C mean
- D uncaring

The words, *cruel*, *mean*, and *uncaring* are all negative choices, and the correct answer is the only positive choice. This keeps the response options from being parallel and makes the answer obvious.

Also, *cruel*, *mean*, and *uncaring* are basically synonyms. A test-wise student who has not read the passage can infer that the correct answer must be the one that means something else. This keeps the response options from being unique.

B. Better

Sarah's mother is **best** described as —

- A wise*
- B talented
- C amusing
- D generous

In this example, all four options are plausible, positive, parallel, and unique.

- 4 Never use “all of the above” or “none of the above” as response options. These are things of the past. Few if any state clients will accept them. The only notable exceptions include choices such as “Correct as it is” on a language test or “Not here” on a mathematics test.

Examples

(Students have read an article titled “The Virginia Dynasty.”)

A. Poor

Which U.S. President was born in Virginia?

- A George Washington
- B Thomas Jefferson
- C James Madison
- D All of the above*

B. Better

Which U.S. President was **not** born in Virginia?

- A George Washington
- B John Adams*
- C Thomas Jefferson
- D James Madison

In Example A, test-wise students who know that options A and B are correct will select the correct response even though they do not know that C is correct. They would receive full credit although they know only two-thirds of the tested content. Students who know the birthplace of only one of the three presidents would get no credit although they know one-third of the tested content. It could be argued that this item has four correct answers.

Example B uses a negative word (**not**) in the stem. As noted elsewhere, this is generally undesirable. However, in this case it is acceptable because the main point is that three of the first four U.S. Presidents were born in Virginia.

More Examples

Poor

Which U.S. President was born in Virginia?

- A George Washington
- B John Adams
- C James Madison
- D A and C but not B*

Although complex response options such as choice D are used on some high-level examinations, they are totally unacceptable on almost all state tests. In addition to the reasons already explained, this type of option favors skillful test takers and discriminates against students who have learned the appropriate content and skills but lack test-taking prowess.

- 5 Avoid trick questions.** Items should be fair, but not overly easy. A mix of easy, moderately difficult, and difficult items is desirable. However, difficulty should come from the content knowledge or thought processes that are required—not from a trick that will trip up students who actually know the content and can apply their knowledge.

Examples

(Students are taking a social studies test.)

A. Poor

Which nation was an ally of the United States during World War II?

- A Japan
- B Germany
- C Russia
- D France*

B. Better

Which nation was an ally of the United States during World War II?

- A Japan
- B Germany
- C Italy
- D Britain*

Example A includes minor nit-picky flaws or tricks. During World War II, the United States was allied with the Soviet Union (not Russia). Also, Germany overran France early in the war, so it is unclear whether the item refers to our French allies who spent much of the war underground or in exile, or to the Vichy government of France, which cooperated with the Germans. In Example B, option D is clearly correct, and A, B, and C are clearly incorrect.

- 6 An item should not offer a cue or clue to its own answer or to that of any neighboring item.** Test-wise students who do not know the content should not be able to match information in the stem with information in the response options to figure out the answer, and they should not be able to use information from one item to correctly answer another. In addition, one item should not depend on another; getting item 10 wrong should not automatically mean getting item 11 wrong.

Examples

(Students are taking a social studies test.)

A. Poor

The invention of the automobile caused Americans to become more —

- A mobile*
- B informed
- C wealthy
- D adventurous

B. Better

The invention of the automobile led directly to an increase in —

- A population
- B wages and prices
- C road construction*
- D leisure time

In Example A, test-wise students can match the word “mobile” with the word “automobile.” This form of cueing/clueing is sometimes called “clang.”

More Examples

(Students read a story about a girl who works hard to make the tennis team.)

A. Poor

Ella is **best** described as —

- 1
- A lucky
 - B talented
 - C average
 - D hardworking*

Why does Ella work so hard?

- 2
- A to impress her friends
 - B to set a good example
 - C to make the tennis team*
 - D to keep herself busy

B. Better

Ella is **best** described as —

- 3
- A lucky
 - B talented
 - C average
 - D hardworking*

What is Ella’s goal?

- 4
- A winning an award
 - B visiting other schools
 - C joining the tennis team*
 - D making new friends

In the Example A, item 2 cues the answer to item 1 for test-wise students. Items 3 and 4 are independent in Example B.

- 7 **Avoid using negative words (e.g., not, none, neither).** Although this rule can sometimes be broken, a negative word needs to be highlighted (e.g., boldfaced, capitalized, and/or italicized depending on the style used in each state). Always avoid double negatives. Especially avoid using negative words in the stem and in the options of the same item.

Examples

A. Poor

Which of the following is **not** unrelated to a decline in the size of Earth's ozone layer?

- A aerosol sprays*
- B smokestack filters
- C replanting in forests
- D continuous monitoring

B. Better

Earth's ozone layer has been damaged by all of these **except** —

- A aerosol sprays
- B automobile emissions
- C replanting in forests*
- D oil well fires

The first stem is an extreme example. Two negatives appear in the stem (i.e., *not*, *unrelated*). The item is confusing even for those who know the content. The second stem is better because there is only one negative. However, using negatives in the stem may not be necessary.

C. Best

Which of the following damages Earth's ozone layer?

- A aerosol sprays*
- B smokestack filters
- C replanting in forests
- D continuous monitoring

Style Issues

Follow the style guidelines set by each state client. There are a number of stylistic decisions and conventions regarding test items. For example:

- When the stem is open, should it end with a dash, a colon, or no mark of punctuation?
- When the stem is open, should each response option end with a period?
- If the stem is closed, but the response options are not complete sentences, should the first letter of each option be in upper or lower case?

There is no one correct style. Therefore, the answers to these style questions depend entirely on the conventions adopted by each client. However, within a state project, consistency is important. Follow *The Chicago Manual of Style* for most detailed style questions, unless otherwise indicated.

Some clients use a format that repeats a sentence from the passage in the item in order to save students the trouble of locating the relevant context. Another style numbers the paragraphs of a passage so that the item stem can refer students to a specific part of the passage.

Be consistent in tone. Present material in a straightforward, factual manner. Avoid a smug, moralistic tone.

Bias and Sensitivity

These are extremely important issues in modern, high-stakes state assessments. The discussion in this manual is by no means the last word. In general, the goal is to avoid topics, language, and allusions that would cause any racial, gender, ethnic, or regional group to be at a disadvantage or to be offended. In addition, equity or “right-to-learn” issues require careful review of all content so that assessments do not favor students of a particular socioeconomic standing or a broader background of experiences.

Bear in mind that students taking these tests may already be apprehensive, and critics of the tests are likely to look for any flaw, no matter how trivial. Therefore, it is important to err on the side of caution. A topic that might be perfectly acceptable in an instructional setting may be inappropriate on a state test. For example, consider a story about the death of a pet. If a student has an emotional reaction in class, the teacher can intervene and excuse the student from the lesson. There is no such opportunity on a state test. If a student is upset, his or her performance on the rest of the test could be adversely affected.

To safeguard against bias, publishers have compiled lists of taboo topics such as the one appearing in the next section.

Topics to Avoid

This guide will help writers identify and avoid subject matter that might be deemed unacceptable for any of the following reasons:

1. The topic is controversial. It might offend teachers, students, or parents. This includes highly controversial topics such as abortion, the death penalty, and evolution. It also includes mildly controversial topics such as smoking.
2. The topic could evoke unpleasant emotions. A student's ability to complete the test could be undermined.
3. The topic shows (or might be perceived to show) bias against a particular group of people.
4. The topic is overly familiar and/or boring to students.

Examples

- Abortion
- Alcohol, including beer and wine
- Behaviors that are inappropriate, including stealing, cheating, lying, and other criminal and/or anti-social behaviors and activities
- Biographies of controversial figures whether or not they are still alive
- Birthdays
- Cancer and other diseases that might be considered fatal (HIV, AIDS)
- Criticism of democracy or capitalism
- Dangerous behavior
- Death of animals or animals dying or being mistreated
- Death, murder, and suicide
- Disasters, including tornadoes, hurricanes, etc. (unless treated as scientific subjects)
- Disrespect of any mainstream racial or religious group
- Double meanings of words that have sexually suggestive meanings
- Evolution
- Family experiences that may be upsetting, including divorce or loss of a job
- Feminist or chauvinistic topics
- Gambling
- Guns and gun control
- Holidays of religious origin (e.g., Halloween, Christmas, Easter)
- Junk food, including candy, gum, chips
- Left- or right-wing politics
- Luxuries (homes with swimming pools, expensive clothes, expensive vacations, and sports activities that typically require the purchase of expensive equipment such as snow skiing)
- Parapsychology
- Physical, emotional, and/or mental abuse, including animal, child, and/or spousal abuse
- Religions (mythology, folk tales, and fables may be problematic also)

- Rock music, including rap and heavy metal
- Sex, including kissing and dating
- Slavery (unless presented in an historical context and presented appropriately)
- Tobacco
- Violence against a particular group of people or animals
- Wars
- Witchcraft, sorcery, or magic
- Words that might be problematic to a specific ethnic group

Exceptions

In certain content areas, sensitive subject matter may be acceptable because it is integral to the course of study. For example, rum, tobacco, slavery, and racial discrimination are topics that are generally avoided in reading passages, even though they represent important, albeit disturbing, events in history. They may be appropriate subject matter on a social studies test that covers content about the triangular trade.

Names

When reading passages are taken from published sources, the characters' names have already been chosen. However, for passages or items that are written specifically for a test, the writer or editor should give careful thought to characters and their names.

To enhance diversity, ethnic names are often desirable. On the other hand, ethnic names are sometimes unfamiliar and difficult to pronounce, especially for poor readers. Good judgment is required to select names that represent diversity without introducing readability problems

Gender Balance

In general, balanced gender diversity is desirable, and women and girls should sometimes (but not always) be depicted performing stereotypically male activities (e.g., playing sports, fixing cars, and building things). Similarly, men and boys should sometimes be depicted cooking, cleaning, and caring for younger children.

Consider the following list of terms and their gender-neutral alternatives. When no proper name is present, gender-neutral terms are always preferred.

<u>Males</u>	<u>Females</u>	<u>Gender Neutral</u>
Actor	Actress	Actor
Chairman	Chairwoman	Chairperson, Chair
Fireman		Firefighter
Mailman		Mail carrier, postal worker, letter carrier
Manhole		Utility hole
Policeman		Police officer
Salesman	Saleswoman	Salesperson
Sportsman	Athlete	Athlete
Waiter	Waitress	Server, waitperson, wait-staff
Fisherman		Fisher, angler, fisherperson

Glossary of Assessment Terms

Anchor (model or exemplar)

An example of a finished student product or written response for a constructed-response item or performance-based task.

Assessment

Gathering evidence to judge a student's demonstration of learning. Assessment aids educational decisions by securing fair, valid, and reliable information to indicate if students have learned what is expected. Assessment is generally built around multiple indicators and sources of evidence (combinations of performances, products, exhibitions, discourse, tests, etc.).

Benchmark

A statement of what students are expected to learn at various developmental levels (e.g., elementary, middle, and high school) to indicate progress made toward meeting a content standard.

Bloom's *Taxonomy of Educational Objectives*

A source used to identify the level of cognitive processing required by an item or activity.

Blueprint

A plan or map that specifies exactly how an assessment is to be designed, which is used throughout the test-development process. It includes a list of the content to be assessed and the numbers and types of items.

Classifying

Grouping entities on the basis of their common attributes.

Comparing/Contrasting

Noting similarities and differences between or among entities.

Competency Test

A test intended to establish whether a student has met minimum standards of skills and knowledge and is thus eligible for promotion, graduation, certification, or other official acknowledgement of achievement.

Comprehending

Generating meaning or understanding.

Content Domain

What the test will measure (e.g., reading comprehension).

Constructed-Response (CR) Item

An item that requires students to produce (construct) a response rather than choosing or selecting an answer option. A constructed-response item might require students to write a sentence or paragraph, or create a chart, diagram, table, map, or timeline. The task must be stated explicitly so students know exactly what is expected.

Criteria

Guidelines, rules, or principles by which students' responses, products, or performances are judged.

Criterion-Referenced Test

A test designed to determine a student's progress toward mastery of a given content area. Items should cover material the student was taught. Performance is compared to an expected level of mastery in a content area rather than to other students' scores. The "criterion" is the standard of performance established as the passing score for the test. These tests are often associated with the phrase "measuring what the student knows and can do," rather than how the test-taker compares to a reference or norm group. Most customized state assessments are CR tests. Superficially, they may look much like Norm-Referenced Tests, but their underlying philosophy is quite different. A criterion-referenced test can have norms, but comparison to a norm is not the purpose of the assessment.

Critical Thinking

Using specific dispositions and skills such as analyzing arguments carefully, seeing points of view, and reaching sound conclusions.

Curriculum

Coherent plan for a designated time period specifying the content knowledge students are expected to understand and apply. A curriculum generally includes standards, benchmarks, and a sequence of content skills that serve as the basis for instruction and assessment.

Cut Score

The score needed to determine the minimum level of performance needed to pass a competency test.

Decision Making

Evaluating and selecting from alternatives.

Distractor (or foil)

Incorrect response option in a multiple choice item.

Elaborate

To analyze, explain, or support a claim by making additional statements.

Evaluating Skills

Core thinking skills that involve assessing the reasonableness and quality of ideas.

Evaluation

Both qualitative and quantitative descriptions of pupil behavior plus value judgments concerning the desirability of that behavior. Using collected information (assessments) to make informed decisions about continued instruction, programs, and activities.

Exemplar

See **Anchor**.

Foil

See **Distractor**.

Grade Equivalent

A score that describes a student's performance in terms of the statistically average student at a given grade level. For example, a grade equivalent score of 5.5 might indicate the student's score could be expected if an average student took the same test in the fifth month of the fifth grade year. This is one of the most misunderstood types of scores. It does not indicate, for example, that a second grader with a grade equivalent score of 5.5 should be promoted to grade 5 or given grade 5 materials.

Identifying Relationships and Patterns

Recognizing ways elements are related.

Inferring

Going beyond available information to reason may be true.

Instruction

The decisions and actions of teachers before, during, and after teaching to increase the probability of student learning.

Mean

One of the measures of central tendency (often called the "average"). Mean is computed by adding all the individual scores and dividing by the number of test subjects in the group. A small number of unusually high or low scores can heavily affect the mean.

Median

Another measure of central tendency, determined by locating the midpoint of all scores. Half are above the median and half are below. A small number of unusually high or low scores will not affect the mean.

Metacognition

The knowledge and awareness of one's own thinking processes and strategies. The ability to consciously reflect on one's own thoughts.

Multiple Choice Item (or Selected Response Item)

An item that contains a question or incomplete statement in the stem and three to four response options or answer choices.

Norm

A distribution of scores obtained from a norm group. The norm is the midpoint (or median) of scores or performances of students in that group. Fifty percent will be above and fifty percent below the norm.

Norm-Referenced Test

A test in which student or group performance is compared to that of a norm group. The results are relative to the performance of an external group and are compared with the norm group providing the performance standard. Often used to measure and compare students, schools, districts, and states on the basis of norm-established scales of achievement. Compare and contrast to criterion-referenced tests.

Observing

An information-gathering skill that involves obtaining information through one or more senses.

Options

The response choices that accompany a question in a selected-response (multiple-choice) format.

Ordering

Sequencing according to a given criterion.

Outcome

An operationally defined educational goal, usually a culminating activity, product, or performance that can be measured.

Percentile

A ranking scale ranging from a low of 1 to a high of 99 with the median score as 50. A percentile rank indicates the percentage of the norm group obtaining scores equal to or less than the test taker's score. A percentile score does *not* refer to the percentage of questions answered correctly.

Performance-Based Assessment

Direct, systematic observation and rating of student performance. For example, a direct writing sample (i.e., an essay test) as opposed to a multiple-choice test with questions about the composing process and the completeness, clarity, and correctness of expression. Proponents of this approach often argue for an ongoing assessment process including features such as portfolios of student work, teacher-student conferences, and student self-reflection. Several state testing programs have attempted to incorporate elements of performance-based assessments, but they have been largely unsuccessful because of the time, expense, legal challenges, and lack of widespread support. Many educators would agree that performance-based assessment is most appropriate in the classroom.

Performance Descriptor

A set of behavioral elements used as a scale to evaluate a student's performance on a criterion-referenced item. Performance descriptors often provide narrative elaboration for the score points on a rubric.

Performance Standards

Descriptive statements of criteria that determine desirable levels of student achievement of content standards central to the curriculum. Performance standards indicate both the nature of the evidence required to demonstrate that the content standards have been met and to rate the quality of the performance.

Performance Task

An assessment item or exercise designed specifically to allow individuals to demonstrate their understanding of content standards.

Portfolio

A systematic and organized collection of a student's work that exhibits to others the direct evidence of a student's efforts, achievements, and progress over time. The collection should involve the student in the selection of its contents, and should include information about the performance criteria, the rubric or criteria for judging merit, and evidence of student self-reflection or self-evaluation. Portfolios may be stored in many formats including written text, electronic text, videos, and physical collections of materials.

Predicting

Anticipating possible outcomes of a situation.

Problem Simulation

A complex assessment activity using a computer. The activity generally requires multiple responses to a challenging question or problem.

Problem Solving

Analyzing and resolving a perplexing or difficult situation.

Process

A method of doing something that generally involves steps or operations that may be ordered or independent. For example, a student engages in the writing process while making notes, outlines, and other organizers prior to writing a draft.

Product

The tangible and stable result of a performance or task. Generally an assessment of student performance is based on evaluation of the product as a demonstration of learning.

Profile

A graphic compilation outlining the performances of an individual on a series of assessments.

Prompt

Information presented in a test item that activates prior knowledge and requires analysis in order for a student to respond. A prompt could be a reading passage, map, chart, graph, drawing, photograph, or combination of these. [Note: Some sources would limit the definition of the word "prompt" to the open-ended question or problem the student must solve, and they would define passages, maps, charts, etc. as "stimuli."]

Quartile

The breakdown of an aggregate of percentile rankings into four categories: the zero to 25th percentile, the 26th to 50th percentile, the 51st to 75th percentile, and the 75th to 99th percentile.

Quintile

The breakdown of an aggregate of percentile rankings into five categories: the zero to 20th percentile; 21st to 40th percentile, etc.

Rating Scale

A scale based on descriptive words or phrases that indicate performance levels. Commonly used terms include minimal, limited, adequate, and proficient.

Recall

A skill that involves retrieving information from memory.

Reliability

The extent to which an assessment yields consistent results. This, along with validity, is a key concept in evaluating the quality of an assessment. Users must have confidence that the same test and parallel forms of the test will yield the same results with repeated administrations.

Rubric

The specific criteria used to determine the caliber of a student's performance. Rubrics may be holistic or item specific depending on the assessment program. See **Scoring Guide**.

Sampling

A way to obtain information about a large group, without testing every member of the group, by examining a small randomly chosen sample that is expected to be reflective of the larger group.

Scale

A classification tool or counting system designed to indicate and measure the degree of to which an event has occurred.

Scale Scores

Scores based on a scale ranging from 001 to 999. Scale scores are useful in comparing performance in one subject area across classes, schools, districts, and other large populations, especially for monitoring changes over time.

Score

A rating or performance based on a scale or classification.

Scoring Guide

A tool for evaluating student performance on an assessment task. It generally includes a set of criteria used to determine the caliber of a student's performance. Different state assessment programs sometimes use the same terms in somewhat different ways. In some states, a "Scoring Guide" is an elaborate booklet that contains rubrics, descriptors for score points, and model papers. A scoring guide developed before items are field-tested might include contrived examples of what student responses are likely to look like. In a later draft, revised after field-testing is complete, these contrived examples may be replaced by authentic student responses.

Selected Response

A type of test item, usually called "multiple choice," that requires students to select a response from a group of possible choices.

Self-Assessment

A process that engages a student in a systematic review of performance. This may involve making comparisons with a standard.

Standard

Statements indicating what students are expected to know and be able to do at a particular grade or upon completing a particular course.

Standardized Test

An objective test that is administered and scored in a uniform manner. Standardized tests may be either norm-referenced or criterion-referenced. They should be constructed carefully and field-tested for appropriateness and difficulty. In most cases, they should be reviewed for bias and sensitivity issues. They are generally accompanied by manuals of directions for administration and score interpretation.

Stem

The item, question, or problem statement.

Strand

A category for classifying the content standards of a subject area curriculum. For example, within the subject area of mathematics, there may be a strand for fractions. Within that strand, there may be more specific benchmarks, objectives, or grade level expectations regarding decimal fractions, improper fractions, etc.

Strategy

A mental process or procedure (or a set of processes or procedures) for problem solving made up of one or more skills. A strategy is usually not a fixed and rigid set of directions. Educators are likely to speak of a process or procedure for simplifying fractions, but they are more likely to speak of a strategy approach to reading in a specific content area (e.g., knowing, among many other things, to attend carefully to text presented in boldfaced type).

Subjective Test

A test in which the assessor's impressions or opinions determine the score or evaluation of a student's performance.

Summarizing

Selecting and combining salient information into a cohesive, concise statement.

Task

A complex assessment activity requiring multiple responses to a challenging question or problem.

Testing

The use of a standardized instrument for the systematic collection of information gathered about a student's knowledge and skills. Standardized tests are just one aspect of a comprehensive system for educational assessment.

Thinking Processes

Relatively complex cognitive operations—such as concept formation, problem solving, and composing—that commonly employ multiple skills.

Validity

The extent to which an assessment measures the desired performance; appropriate inferences can be concluded from these results. Along with reliability, validity is a key concept in evaluating the quality of an assessment. Users must have confidence that the assessment accurately reflects the learning it was designed to measure.

Science DOK Levels

Please note that, in science, “knowledge” can refer both to content knowledge and knowledge of scientific processes. This meaning of knowledge is consistent with the *National Science Education Standards* (NSES), which terms “Science as Inquiry” as its first Content Standard.

Level 1 (Recall and Reproduction) requires the recall of information, such as a fact, definition, term, or a simple procedure, as well as performance of a simple science process or procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. A “simple” procedure is well defined and typically involves only one step. Verbs such as “identify,” “recall,” “recognize,” “use,” “calculate,” and “measure” generally represent cognitive work at the recall and reproduction level. Simple word problems that can be directly translated into and solved by a formula are considered Level 1. Verbs such as “describe” and “explain” could be classified at different DOK levels, depending on the complexity of what is to be described and explained.

A student answering a Level 1 item either knows the answer or does not: that is, the item does not need to be “figured out” or “solved.” In other words, if the knowledge necessary to answer an item automatically provides the answer to it, then the item is at Level 1. If the knowledge needed to answer the item is not automatically provided in the stem, the item is at least at Level 2. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Recall or recognize a fact, term, or property.
- Represent in words or diagrams a scientific concept or relationship.
- Provide or recognize a standard scientific representation for simple phenomenon.
- Perform a routine procedure, such as measuring length.

Level 2 (Skills and Concepts) includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is **more complex** than in Level 1. Items require students to make some decisions as to how to approach the question or problem. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply **more than one step**. For example, to compare data requires first identifying characteristics of the objects or phenomena and then grouping or ordering the objects. Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts. Some action verbs, such as “explain,” “describe,” or “interpret,” could be classified at different DOK levels, depending on the complexity of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is a Level 2. An item that requires interpretation from a complex graph, such as making decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated, is at Level 3. Some examples that represent, but do not constitute all of, Level 2 performance, are:

- Specify and explain the relationship between facts, terms, properties, or variables.
- Describe and explain examples and non-examples of science concepts.
- Select a procedure according to specified criteria and perform it.
- Formulate a routine problem, given data and conditions.
- Organize, represent, and interpret data.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity does not result only from the fact that there could be multiple answers, a possibility for both Levels 1 and 2, but because the multi-step task requires more demanding reasoning. In most instances, requiring students to explain their thinking is at Level 3; requiring a very simple explanation or a word or two should be at Level 2. An activity that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Experimental designs in Level 3 typically involve more than one dependent variable. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems. Some examples that represent, but do not constitute all of Level 3 performance, are:

- Identify research questions and design investigations for a scientific problem.
- Solve non-routine problems.
- Develop a scientific model for a complex situation.
- Form conclusions from experimental data.

Level 4 (Extended Thinking) involves high cognitive demands and complexity. Students are required to make several connections—relate ideas within the content area or among content areas—and have to select or devise one approach among many alternatives to solve the problem. Many on-demand assessment instruments will not include any assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. “Develop generalizations of the results obtained and the strategies used and apply them to new problem situations,” is an example of a grade 8 objective that is a Level 4. Many, but not all, performance assessments and open-ended assessment activities requiring significant thought will be Level 4.

Level 4 requires complex reasoning, experimental design and planning, and probably will require an extended period of time either for the science investigation required by an objective, or for carrying out the multiple steps of an assessment item. However, the extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be a Level 4. Some examples that represent, but do not constitute all of, a Level 4 performance are:

- Based on data provided from a complex experiment that is novel to the student, deduct the fundamental relationship between several controlled variables.
- Conduct an investigation, from specifying a problem to designing and carrying out an experiment, to analyzing its data and forming conclusions.

Examples Applied to Science Objectives and Assessment Items

Sample Science Objectives

Use the science DOK levels on the previous pages to determine the DOK levels for the following five sample objectives. Except for the last, these objectives are for grade 8. When you are finished, turn the page to see whether you agree with the way we coded these objectives! Then try using the DOK levels on the 10 sample science items in Part ii.

Objective 1. Students should identify the structure and function of the major parts of animal and plant cells.

Objective 2. Students should design and conduct a science investigation in their home or community that involves data collection, display, and interpretation.

Objective 3. All students will analyze claims for their scientific merit and explain how scientists decide what constitutes scientific knowledge; show how science is related to other ways of knowing; show how science and technology affect our society; and show how people of diverse cultures have contributed to and influenced developments in science.

Objective 4. All students will measure and describe the things around us; explain what the world around us is made of; identify and describe forms of energy; and explain how electricity and magnetism interact with matter.

Objective 5. (Grade 10) Students should be able to explain the process of photosynthesis in detail.

DOK Levels of the Sample Science Objectives

Objective 1. Level 1. “Identifying” the cell parts and their functions only involves recalling and naming/labeling.

Objective 2. Level 4. This requires extended time and involves all of the major aspects of a scientific investigation. If the most involved type of activity that a scientist ever engages in is not a Level 4 activity, then what is?

Objective 3. Level 3. The activities described in this objective require synthesis of different kinds of information, analysis of information, and criticism based on scientific methodology, and deep explanation.

Objective 4. Level 2. It is difficult to determine the DOK level for an objective with many parts like this. Measuring and identifying are typically Level 1 activities, but describing and explaining can signify different levels. With the exception of the last phrase of this objective, the descriptions and explanations asked for here are of *things* rather than *processes*, explanations of *what* rather than *how*. However, “explain how electricity and magnetism interact with matter” could call for some synthesis of different kinds of information, which would signify a higher level of knowledge. On the other hand, the explanation asked for here could be quite simple, too. So parts of this objective are Level 1 and parts are Level 2. What should we do? In such a case, you should code the objective according to the *highest* depth of knowledge that it requires the student to display, even if this DOK level is only found in one part of the objective.

Objective 5. Level 2. Students here not only must recall simple definitions and terms, but must also be able to describe and explain a process. On the other hand, this does not require any strategic reasoning, such as using the process of photosynthesis to make sense of an observed phenomenon.

Sample Science Assessment Items

Now try coding some sample assessment items using the science DOK levels. There are six items for grade 8 and four for high school. After you are finished coding these, read our “answers” on the following page.

The following six items are from grade 8 assessments:

1)

Which group of organisms would all be found living in a tropical rain forest?

- A) Lizards, insects, cacti, kangaroos
- B) Vines, palm trees, tree frogs, monkeys
- C) Evergreens, moose, weasels, mink
- D) Lichens, mosses, caribou, polar bears

2) Make a graph of your heart rate as you walk in place for five minutes.

3)¹

The purpose of this task is to determine where, how high, and for what purpose (flood control, recreation, hydroelectric power, etc.) to build a dam. You will have a total of 45 minutes to complete this task. You may use up to 20 minutes to complete the group work, found on the first two pages of this form. When you finish the group activity, someone from your group should tell the facilitator. Then you may open this form and follow the directions inside by yourself.

Your group should have the following materials:

- Plastic model
- Clay
- Water in a pitcher
- Map
- Ruler
- Paper towels

Pencils (cont'd on next page)

GROUP ACTIVITY (cont'd from previous page)

1. Examine the model of the river valley as well as the map you have been provided. Using this information, discuss possible sites for a dam.
2. Use the clay to construct a dam on the model. With the water, test the impact of your dam on the nearby areas. Try different locations and dam heights based upon the dam's purpose. Record the different locations on the group's map. Record information from the trials in the chart on the next page.

Record information from your group's tests in this chart.

Site #	Location	Purpose	Impact

¹ [This item was contributed to the PALS (Performance Assessment Links in Science) website (<http://www.ctl.sri.com/pals/>) by the Kentucky Department of Education.]

Make sure that each group member's name appears on the map. One member of the group should insert the map into his or her response form when passing in the completed form.

When you are finished with the work on this page, one member of the group should tell the facilitator that your group has finished its group work. Then go on to the individual work. Remember that you must work alone on those pages. You may not discuss the questions or share information.

INDIVIDUAL ACTIVITY

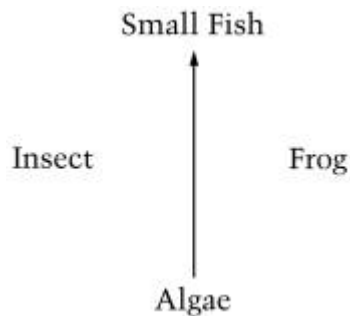
3. After reviewing the work your group has done, where would you place the dam and how high would you make it? Why?
4. What social, environmental, and economic impacts would the location you chose for the dam have on the surrounding community?
5. Describe concerns you would include in an environmental impact statement for dam sites other than the one you selected in question 3.

Be sure one member of the group inserts the map inside his or her form for collection.

4) When operating, ordinary incandescent lightbulbs produce a lot of heat in addition to light. Fluorescent lightbulbs produce much less heat when operating. If you wanted to conserve electricity, which type of bulb should you use? Explain your answer.

5)

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.)



6)

Suppose that a farmer near the pond sprayed crops with a pesticide to kill insects and that some of the spray washed into the pond. (This pesticide breaks down very slowly.) If several months later a biologist tested all the organisms in the pond system for the pesticide, which organism would most likely have the greatest concentration of the pesticide? Explain your answer.

The following six items are from High School assessments. The first two refer to this passage:

During the development of chemistry, many chemists attempted to explain the changes that occur when *combustible* (capable of burning) materials burn and metals corrode or rust. The following are two proposed theories.

Phlogiston Theory

According to this theory, combustible materials, such as wood, coal, or metal contain a massless "essence" or presence called phlogiston. When combustion occurs, the phlogiston is released from the combusting object and is absorbed by the air. For example, when a piece of wood is burned, phlogiston is released to the air and the wood is converted to ash. The ash is free of phlogiston and can no longer support combustion. Similarly, if a metal is heated, the phlogiston is lost to the air and the metal is converted into a nonmetallic, powdery substance called ash, or calx. The *corrosion* (changing of a substance by a chemical reaction) of metals, such as the rusting of iron (Fe), also involves the loss of phlogiston from the metal, but at a slower rate than burning. Rust can be turned back into metal by heating it in air with a substance rich in phlogiston, such as charcoal. A transfer of phlogiston from the charcoal to the rust converts the rust back to metal.

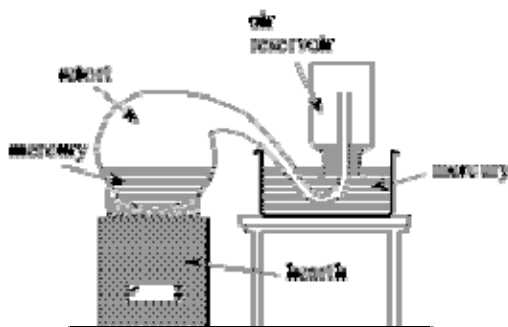
Oxygen Theory

According to this theory, burning and rusting involve an element called oxygen, which is found in the air. The complete combustion of a piece of wood involves the rapid reaction of the wood with oxygen gas (O_2) to produce carbon dioxide (CO_2), which is a nonflammable gas, and water (H_2O). The rusting of iron involves the slow reaction of iron with oxygen to produce iron oxides such as Fe_2O_3 . These iron oxides are known as rust. Heating rust with charcoal produces iron because the charcoal combines with the oxygen in the rust. In these transformations, there is a *conservation of mass* (the total mass of the reactants must equal the total mass of the products in a chemical reaction). In these reactions matter is neither created nor destroyed, but merely transformed.

7) According to the Phlogiston Theory, the gases collected from the complete burning of a piece of charcoal in air would be capable of:

- F. converting the ash from corroded tin back to tin metal.
- G. supporting combustion of another piece of charcoal.
- H. rusting iron.
- J. converting wood ash into rust.

- 8) A chemist heated a sample of mercury for several days in the apparatus shown below. As the experiment proceeded, the mercury in the retort became covered with a red powder, and the volume of mercury increased in the air reservoir. The remaining material in the reservoir would not support combustion. Which of the following theories is supported by the results of this experiment?



- A. The Phlogiston Theory, because the red powder resembled an ash
- B. The Phlogiston Theory, because the air in the reservoir could not support combustion and therefore did not contain oxygen
- C. The Oxygen Theory, because the mercury level dropped in the air reservoir indicating increased oxygen content
- D. The Oxygen Theory, because the mercury level rose in the air reservoir indicating decreased oxygen content

The following sample high school assessment items do not use the above passages.

- 9) A scientist synthesizes a new drug. She wants to test its effectiveness in stopping the growth of cancerous tumors. She decides to conduct a series of experiments on laboratory mice to test her hypothesis.

What should she do?

- a. Give half the mice the drug, the other half none, and compare their tumor rates.
- b. Give the drug to all mice, but only to half every other day, and record tumor rates.
- c. Double the dosage to all mice each day until tumors start to disappear.
- d. Give the drug only to those mice who have tumors and record their weights.

10) The results of one of her experiments are shown in the table below:

Average tumor size in millimeters by dosage and days of treatment							
Dosage	Days of Treatment						
	1	7	14	21	28	35	42
150mg	5	6	8	11	13	15	18
300mg	5	5	6	7	7	9	10
600mg	5	5	4	4	5	4	3

What can she conclude from these results?

- a. The effectiveness of the drug over time depends on the size of the dosage.
- b. The drug is effective over time regardless of the size of the dosage.
- c. The size of the dosage affects tumor size regardless of the length of time.
- d. The drug is ineffective regardless of the dosage or length of time.

11) What is the process called which plants use to manufacture sugar from sunlight?

12) In a laboratory experiment using spectrophotometry, an enzyme is combined with its substrate at time zero. The absorbance of the resulting solution is measured at five-minute intervals. In this procedure, an increase in absorbance is related to the amount of product formed during the reaction. The experiment is conducted using three preparations as shown in the table below.

Enzyme preparation	Absorbance				
	0 min	5 min	10 min	15 min	20 min
I. 3 mL enzyme, 2 mL substrate, pH 5	0.0	0.22	0.33	0.38	0.37
II. 3 mL boiled enzyme, 2 mL substrate, pH 5	0.0	0.06	0.04	0.03	0.04
III. 3 mL enzyme, 2 mL substrate, pH 6	0.0	0.32	0.37	0.36	0.38

The most likely reason for the failure of the absorbance to increase significantly after 10 minutes in preparation III is that

- the reaction is thermodynamically impossible at pH 6
- the enzyme is not active at this pH
- a pH of 6 prevents color development beyond an absorbance of 0.38
- the enzyme is degraded more rapidly at pH 6 than it is at pH 5
- most of the substrate was digested during the first 10 minutes

DOK Levels for the Science Sample Assessment Items

Grade 8 Items:

- 1) Level 1. This item assesses “the recall of information such as a fact or definition.”
- 2) Level 2. This item has several steps and requires some decision making. Students must decide appropriate intervals for measuring pulse and procedures for graphing data. “Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.”
- 3) Level 4. An example in the Level 4 definition is “Conduct an investigation, from specifying a problem to designing and carrying out an experiment, to analyzing its data and forming conclusions.” This item requires students to perform the breadth of activities an actual scientist would perform and demands extended time and thought.
- 4) Level 3. If this did not require an explanation, it would be Level 1. But here students must explain the complex connection between electrical consumption and production of heat in order to receive full credit. “In most instances, requiring students to explain their thinking is at Level 3.”
- 5) Level 1. Even though this item has multiple steps, the steps are not interrelated and do not increase the item’s cognitive demands. Each step involves only recall.
- 6) Level 3. Explaining a simple and short answer can be Level 2, but the explanation required here is much more involved. The rubric requires giving full credit only if the student response “names the highest animal on the food chain, the heron, as having the greatest concentration of the pesticide.” In addition, the response must demonstrate an understanding of biological magnification by explaining that the heron accumulates the greatest concentration of the pesticide from the fish it eats because the fish have accumulated the pesticides from the organisms they have eaten.”

High School Items:

- 7) Level 3. Although it is uncommon, it is possible for a multiple choice item to be at Level 3. This item employs demanding reasoning, because it requires the student to make a complex inference based on an unfamiliar theory.
- 8) Level 3. Like the previous item, this involves making complex inferences from two conflicting theories. This non-routine problem also requires “drawing conclusions from observations” and “explaining phenomena in terms of concepts.”

9) Level 2. Students must at least apply knowledge of controlled-experiment design to this situation, or derive it from the choices offered.

10) Level 2. If this item was open-ended, asking what conclusions could be drawn from the data and why, then it would be Level 3. Here the student only needs to check which of the presented solutions is most reasonable, which requires no decision-making or creativity.

11) Level 1.

12) Level 3. This is another example of a multiple-choice item that is still Level 3, this time due to the complexity of the presented situation. Students must compare the interaction of two dependent variables and interpret the data in light of a complex body of interrelated concepts.

Comparison of Bloom's Taxonomy and Webb's Depth of Knowledge

Bloom's Taxonomy

Level	Categories	Abbreviated Definition	Possible Science Action Words*
1	Knowledge	student remembers, recalls appropriate previously learned information	identify, recall, observe, recognize, use, calculate, measure, order
2	Comprehension	student translates, comprehends, or interprets information based on prior learning	explain, interpret, describe, classify, identify, recognize, predict
3	Application	student selects, transfers, and uses data and principles to complete a task or problem with a minimum of direction	apply, classify, experiment, interpret, use, order, calculate
4	Analysis	student distinguishes, classifies, and relates the assumptions, hypotheses, evidence, or structure of a statement or question	analyze, order, explain, classify, arrange, compare, contrast, infer, calculate, categorize, examine, experiment, question, test
5	Synthesis	student originates, integrates, and combines ideas into a product, plan or proposal that is new to him or her	combine, arrange, rearrange, modify, invent, design, construct, organize, predict, infer, conclude, create experiment and record data
6	Evaluation	student appraises, assesses, or critiques on a basis of specific standards and criteria	evaluate, measure, explain, compare, summarize, predict, test, decide, rate, conclude

Webb's Depth of Knowledge

Level	Categories	Abbreviated Definition	Possible Science Action Words*
1	Recall	student recalls facts, information, and definitions; can perform a routine procedure	identify, recall, observe, recognize, use, calculate, measure, order
2	Basic Application of Skill / Concept	student uses information, conceptual knowledge, and procedures; demonstrates the relationship between facts, terms, properties or variables; organizes, represents and interprets data	explain, interpret, describe, classify, identify, order, recognize, predict, apply, use, calculate, organize, estimate, observe, collect and display data
3	Strategic Thinking	student uses reasoning; develops a plan or sequence of steps, draws conclusions from experimental data and observations, solves non-routine problems	analyze, order, explain, classify, arrange, compare, contrast, infer, interpret, calculate, categorize, examine, experiment, question, predict, evaluate, test
4	Extended Thinking	student conducts an investigation needs time to think and process multiple conditions of a problem or task, develops generalizations	combine, arrange, rearrange, propose, evaluate modify, invent, design, construct, organize, predict, infer, conclude, evaluate, create experiment and record data

*Some action verbs can be classified at different depth-of-knowledge levels depending on the context of the item and the complexity of the action.

Fairness in Testing



Guidelines for Training Bias, Fairness, and Sensitivity Issues

Table of Contents

Introduction	3
Definition of Bias	4
Types of Bias	5
Stereotyping	5
Gender Bias	7
Regional or Geographical Bias	9
Ethnic or Cultural Bias	9
Socioeconomic or Class Bias	10
Religious Bias	10
Ageism (Bias Against a Particular Age Group)	11
Bias Against Persons with Disabilities	11
Experiential Bias	12
Maintaining Balance	13
Topics to Avoid	14
Special Circumstances	15
Historical Contexts	15
Literary Contexts	15
Points to Remember	16
Sample Review Form	17
References	18
Sample Items with Bias, Fairness, and/or Sensitivity Concerns	19

Introduction

The most important part of the development of any new test is to ensure balanced treatment and control of potential bias, stereotyping, and insensitivity in the items or in the test-related materials. Data Recognition Corporation (DRC) understands that the presence of any type of bias in a test is undesirable not only from a civil rights point of view, but also from a measurement point of view. Issues of bias, fairness, and sensitivity in testing can have a direct impact on test scores. Our test developers are committed to the development of items and tests that are fair for all students. At every stage of the item and test development process, we employ procedures that are designed to ensure that our items and tests meet Standard 7.4 of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999).

Standard 7.4: Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.

In meeting Standard 7.4, DRC employs a series of internal quality steps that we believe are among some of the best in the industry. We provide specific training for our test developers, item writers, and reviewers on how to write, review, revise, and edit items for issues of bias, fairness, and sensitivity, as well as for technical quality. Our training also includes an awareness of and sensitivity to issues of cultural diversity.

In addition to providing *internal* training in reviewing items in order to eliminate potential bias, we also provide *external* training to our clients, including state departments of education, review panels of minority experts, teachers, and other stakeholders. DRC understands the importance of having external panels with a wide variety of expertise in reviewing items and tests for potential bias. External panels of professionals provide a review of items for subtle forms of bias that often can be perceived only by individuals who possess a wide variety of appropriate expertise and represent specific constituencies.

This manual has been prepared to summarize DRC's guidelines for bias, fairness, and sensitivity, including how to eliminate language, symbols, words, phrases, and content that might be considered offensive by members of racial, ethnic, gender, or other groups. Our guidelines may be modified to meet client's requirements and/or state-specific guidelines.

Definition of Bias

While there are many definitions of bias, the following definition is provided on page 76 of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999):

The term *bias* in tests and testing refers to construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees. In other words, **bias is the presence of some characteristic of an item and/or test that results in two individuals of the same ability but from different subgroups performing differently on the item and/or test.** Therefore, it is most important that there are no ambiguities in the test items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries.

Types of Bias

There are many types of bias. They include stereotyping and discriminating against people because of gender, regional or geographical differences, ethnicity or culture, socioeconomic or class status, religion, or age, as well as bias against other groups of people, including those with disabilities. Another form of bias involves the use of questions and/or activities in the items or on a test as a whole that are not relevant to the life experiences of the students responding to the items or test. A definition of each type of bias, along with samples, is provided below.

Stereotyping

“Stereotype is an image formed by ascribing certain characteristics (e.g., physical, cultural, personal, occupational, historical) to all members of a group” (National Evaluation Systems, Inc. page 2). Stereotyping in test items and tests might include physical characteristics, intellectual characteristics, emotions, careers, activities, and domestic or social roles. In writing or reviewing test items, it is very important that all groups are portrayed fairly, without stereotyping. As a result, there should be a range of characteristics, careers, and social roles across all groups, and no one group should be characterized by any one particular attribute or characteristic. Following are examples of stereotyping.

Stereotype

Examples

Physical characteristics

Males are strong and capable leaders.
Females are weak.
The elderly are feeble and sickly.
Children are healthy and full of energy.
The elderly are dependent upon others.
People with disabilities are dependent upon others.
Females worry about their hair.

Intellectual characteristics

Males do better in mathematics and science.
Females do better in reading and language arts.
Asian Americans excel in academics.

Emotions

Males are aggressive, courageous, and strong.
Females are weak, weepy, tender, and fearful.

Types of Bias
Stereotyping (continued)

Stereotyping

Examples

Careers

Females are nurses, teachers, and secretaries.
Males are doctors, principals, superintendents,
lawyers, and skilled laborers (e.g., plumbers,
construction workers, painters).
African-Americans are athletes.
Hispanics operate lawn care businesses.
Asian-Americans own dry cleaning businesses.

Activities

Females play with dolls and read books.
Females do domestic chores (e.g., clean house,
cook, sew).
Females spend money.
Males play sports and work with tools.
Boys are rowdy.
Girls are quiet.

Domestic and/or Social Roles

Females are responsible for childcare.
Men work outside of the home and are the
breadwinners.

Community

Asian-Americans live in ethnic neighborhoods.
African-Americans live in high-rise apartment
buildings located in urban areas.
American Indians live on reservations.

Leadership

Men are leaders and rulers.
Women are followers.
Women are dependent on men.
Men are elected to political positions.
Females in leadership roles are aggressive and
pushy.

Types of Bias (continued)

Gender Bias

Gender bias involves items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries that show members of either sex in stereotypical activities, emotions, occupations, characteristics, and/or situations. Gender bias also involves the use of demeaning labels.

Examples of gender bias

Titles and specific terms referring to humanity at large, such as

- Mankind
- Manhood
- Manpower
- Man of the hour
- Man-hours
- Man-made

Use of gender specific terms for occupations, such as

- Fireman
- Workman
- Chairman
- Policeman
- Mailman
- Salesman
- Insurance man
- Businessman
- Congressman

Use of pronouns that imply a stereotype, such as

- The nurse went to the hospital, and *she* was able to talk with the patient.
- The factory worker needed to earn more money for *his* family.
- When the lawyer delivered *his* closing remarks, the jury listened carefully.
- A politician must give a lot of speeches when *he* runs for office.

Use of phrases that identify genders in terms of their roles or occupations, such as

- Men and girls were invited to the lecture.
- The travelers took their wives and children with them.
- The happy couple was introduced as man and wife.

Types of Bias

Gender Bias (continued)

Use of phrases or words with an emphasis on marital status, such as

- Abraham Lincoln and Mrs. Lincoln attended the play.
- George Washington and Martha visited the new building.
- Dr. and Mrs. Jones attended the opening of the new warehouse.
- The admirable Dr. George Halstead and his wife, Maria, visited the library.

Use of words that identify genders in the salutation of a business letter, such as

- Dear Sir:
- Dear Madam:
- Dear Gentlemen:

Use of words or phrases that are not parallel, such as

- The girls' restroom is down the hall, and the men's restroom is on the second floor.
- The boys' locker room door is painted green, and the women's locker room door is painted yellow.
- The men's department is on the right; the ladies' department is on the left.

Use of figures of speech, such as

- Old wives' tale
- Right-hand man
- Man versus nature
- The best man for the job
- The better half

Use of gender-specific terms or diminutive words, such as

- Sweet young thing
- Usherette
- Housewife
- Maid
- Cleaning lady
- Little woman
- Career girl
- Houseboy
- Steward

Types of Bias (continued)

Regional or Geographical Bias

Regional and/or geographical bias involves items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries that include terms that are not commonly used nationwide or within a particular region or state to which the test will be given. It also involves the use of terms that have different connotations in different parts of the country and/or geographical regions. It is important to note that some experiences may not be common to all students. For example, within a given geographic area not all students might be familiar with snow, so questions involving sleds and toboggans, for example, may well reflect a regional or geographical bias.

Examples of regional or geographical bias

- She ordered a new davenport (couch or sofa).
- Go get your toboggan (hat or type of sled).
- The students stood in line at the bubbler (water fountain or drinking fountain).
- Turn left at the berm (curb).
- Take the pike (road).

Ethnic or Cultural Bias

Ethnic bias involves items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries that include terms that are demeaning and/or offensive to a particular ethnic group or culture. In addition, no minority group should be portrayed as being uneducated or poor.

Examples of ethnic or cultural bias

- Maria was in the kitchen making tacos.
- The Chinese owned a laundry in our area.
- Native Americans are very close to nature.

Terminology

Terms that have a negative connotation or that reinforce negative judgments should also be avoided. Following is a list of **acceptable** terms.

- African-American
- Asian-American or Pacific Island American
- Latino, Mexican-American, Hispanic
- Tribal name (preferred), Native American, American Indian
- European-American

Types of Bias (continued)

Socioeconomic or Class Bias

Socioeconomic or class bias involves items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries that include activities, possessions, or ideas that may not be common to all students within a given area. For example, not all students in a given area own CD players or video games, nor do all students in a given area participate in certain sports activities, such as golf, snow skiing, or sailing. In addition, not all students in a given area take expensive vacations or attend expensive schools.

Examples of socioeconomic or class bias

- They were members of the country club.
- Boarding school.
- How many golf balls landed in the lake?
- The club members plan to go snow skiing over the holidays.
- My great aunt lives in a town house overlooking Lake Michigan.

Religious Bias

Religious bias involves items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries that include terms that are demeaning and/or offensive to a particular religious group.

Examples of religious bias

- The house on Smith Street is decorated for Halloween.
- There were several Christmas trees in the window.
- The students in the class will stand and say the *Pledge of Allegiance*.
- The high school students will be attending a rock-and-roll dance at the community center.

It is also important to note that no religious belief or practice should be portrayed as a universal norm or as inferior or superior to any other.

Types of Bias

Ageism (Bias Against a Particular Age Group)

There are other subtle forms of bias, including bias against the elderly or ageism. Ageism involves items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries that include terms that are demeaning and/or offensive to the elderly or older persons (65 years or older). Ageism can also involve issues of bias with other age groups, including teenagers and young children.

It is important to note, however, that representing older persons or any age group fairly does not mean that the content of the items has to be revised or rewritten to seem unrealistic. Rather, as a whole, the items and the test should show older people or any age group in a variety of roles and activities whenever they appear naturally in the test content.

Examples of ageism (bias against a particular age group)

- Despite the fact that she was very old, she was able to walk down the stairs.
- The child's grandfather seemed senile.
- They were acting like typical irresponsible teenagers.

Bias Against Persons with Disabilities

Another form of subtle bias involves issues of bias related to persons with disabilities. This type of bias involves items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries that include terms that are demeaning and/or offensive to persons with disabilities. It is important to note, however, that representing persons with disabilities does not mean that the content of the items has to be revised or rewritten to seem unrealistic. Rather, as a whole, the items and the test should show people with disabilities in a variety of roles and activities whenever they appear naturally in the test content.

Examples of bias against persons with disabilities

- After the car accident, the student was confined to a wheelchair.
- He became a successful writer despite his disability.
- She is a blind person.
- The student is handicapped.
- The child made great strides in overcoming her disability.

Types of Bias

Bias Against Persons with Disabilities (continued)

Terminology

Terms that have a negative connotation or that reinforce negative judgments (crippled, victim, afflicted, confined, etc.) should also be avoided. It is also important that no one with a disability should be pictured as helpless or portrayed as pitiful.

Do not use

Retarded
Hard of hearing
Deaf and Dumb or Deaf-mute
Learning-disabled
Handicap

Use

Developmentally delayed
Hearing impaired
Deaf or hard-of-hearing used accurately
Person with a learning disability
Disability
Visually-impaired or Blind used accurately

Experiential Bias

The questions and activities reflected in the items or test, as a whole, should also be relevant to the life experiences of the students responding to the items. In other words, for a student to respond sensibly to the test questions, he or she must know what the question is about. In addition, culturally specific knowledge should be avoided, along with the use of difficult words and figures of speech.

Examples of experiential bias

- Pat knew she would win the race as she had an ace up her sleeve.
- Put the pedal to the metal and clean up your room.
- I needed change for the subway turnstile.
- The arroyos filled quickly during the storm.
- The super takes care of cleaning the foyer.

Maintaining Balance

Bias may also occur as a result of having a lack of balance through underrepresentation of a particular ethnic group and/or gender. Therefore, whenever possible, tests and test-related materials should contain content that is balanced across ethnic groups and across gender. The content of the pool of items and/or test, as a whole, should also reflect cultural diversity. In order to achieve balance, the test developers at DRC review the pool of items or the test, as a whole, to determine whether or not there is an adequate representation of

- Females and males in both traditional and nontraditional roles
- Female and male names
- Minority groups in various environments and occupations
- Minority groups, including the use of names

The issue of fairness also involves content inclusiveness. Subtle forms of bias can result from omitting certain areas of information and/or from omitting certain topics. Wherever possible, the content should show people in everyday situations and groups should be depicted as fully integrated in the society, reflecting the diverse multicultural composition of society as a whole (NES, page 9).

Topics to Avoid

Because issues of bias, fairness, and sensitivity in testing can have a direct impact on the test scores, it is also important that sensitive and offensive topics be avoided. A topic might be considered offensive or controversial if it offends teachers, students, parents, or the community at large. This includes highly charged and controversial topics such as abortion, the death penalty, and evolution. Unacceptable content might also include less controversial topics, such as the use of tobacco or topics that could evoke unpleasant emotions on the part of a given student. In addition, topics that appear to promote or defend a particular set of values should be avoided. It is important to remember that the ability of the student to take the test should never be undermined. Following are examples of topics generally to be avoided.

Examples of topics to be generally avoided

- Abortion
- Alcohol, including beer and wine
- Behaviors that are inappropriate, including stealing, cheating, lying, and other criminal and/or anti-social behaviors and activities
- Biographies of controversial figures whether or not they are still alive
- Birthdays
- Cancer and other diseases that might be considered fatal (HIV, AIDS)
- Criticism of democracy or capitalism
- Dangerous behavior
- Death of animals or animals dying or being mistreated
- Death, murder, and suicide
- Disasters, including tornadoes, hurricanes, etc. (unless treated as scientific subjects)
- Disrespect of any mainstream racial or religious group
- Double meanings of words that have sexually suggestive meanings
- Evolution
- Family experiences that may be upsetting, including divorce or loss of a job
- Feminist or chauvinistic topics
- Gambling
- Guns and gun control
- Holidays of religious origin (e.g., Halloween, Christmas, Easter)
- Junk food, including candy, gum, chips
- Left- or right-wing politics
- Luxuries (homes with swimming pools, expensive clothes, expensive vacations, and sports activities that typically require the purchase of expensive equipment such as snow skiing)
- Parapsychology
- Physical, emotional, and/or mental abuse, including animal, child, and/or spousal abuse

- Religions, except in appropriate historical context; mythology, folk tales, and fables may contain religious elements as part of appropriately presented literary excerpts.
- Sex, including kissing and dating
- Slavery (unless presented in an historical context and presented appropriately)
- Tobacco
- Violence against a particular group of people or animals
- Rock music, including rap and heavy metal
- Wars
- Witchcraft, sorcery, or magic
- Words that might be problematic to a specific ethnic group

Special Circumstances

In certain subject areas, a sensitive topic may be acceptable because the topic is a part of the course of study or may be required in order to measure the specific curriculum content standards and/or test objectives. For example, it may be appropriate to have test questions dealing with hurricanes. However, the questions should not focus unduly upon the destruction of property or the deaths of human beings. Other special circumstances include historical and literary contexts. A discussion of these special circumstances is provided below.

Historical Contexts

In order to measure the content curriculum standards, social studies tests often include topics that might otherwise be deemed as controversial. For example, in a history test, the topic of slavery might be used. The student would know that such a controversial topic is used to access knowledge of a particular curriculum content standard and/or set of objectives and, therefore, the topic would not reflect the views of the test developer.

Literary Contexts

Today's tests often require the use of authentic or previously published passages. As a result, sometimes a given passage or prompt might contain controversial material, including sentences, phrases, and/or words. If the overall passage or prompt is acceptable, it may be possible to edit and or delete the objectionable sentences, phrases, words, and/or references in order to eliminate the potential bias. In such cases, DRC test developers request permission from the publisher to make such edits and/or changes, and they would do so only if permission is granted.

Points to Remember

When reviewing items (questions and responses), passages prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries for issues of bias, fairness, and sensitivity, the following questions should be asked.

1. Do the items (questions and responses), passages, prompts, stimulus materials, artwork, graphs, charts, and test-related ancillaries:

DemEAN any religious, ethnic, cultural, or social group?

Portray anyone or any group in a stereotypical manner?

Contain any other forms of bias, including gender, regional or geographical, ethnic or cultural, socioeconomic or class, religious, age-related bias, or bias against persons with disabilities?

2. Are there any topics that might disadvantage a student for any reason?
3. Are there any culturally specific sets of knowledge, terms, difficult words and/or figures of speech that might disadvantage a group of students?
4. Are the questions and activities reflected in the items or test, as a whole, relevant to the life experiences of the students responding to the items?
5. As a whole, does the test or pool of items have a balance across ethnic groups and across genders, including an adequate representation of:

Females and males in both traditional and nontraditional roles

Female and male names

Minority groups in various environments and occupations

Minority groups, including the use of ethnic names

6. Wherever possible, does the content show minority groups in everyday situations and groups depicted as fully integrated in the society, reflecting the multicultural composition of society as a whole?

Sample Review Form

Name: _____

Date: _____

Subject Area: _____ **Grade Level:** _____

[illegible]

Comments:

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Haladyna, T. (1999). *Developing and validating multiple-choice test questions*. Mahwah, New Jersey: Lawrence Erlbawn.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*. Washington DC.

McDivitt, P.J., Newsome, D., Shoffner, M., Wall, J., and Watts, R. (2002). *Applying the standards for educational and psychological testing: what teachers and counselors need to know*. Association for Assessment in Counseling.

National Evaluation Systems, Inc. (1990). *Bias concerns in test development*. Washington DC: The National Evaluation Systems, Inc. (NES).

Osterlind, S.J. (1998). *Constructing test items: multiple-choice, constructed-response, performance, and other formats*, second edition. AH Dordrecht The Netherlands: Kluwer Academic Publishers.

Sandoval, J., Frisy, C.L., Geisinger, K.F., Scheuneman, J.D., and Grenier, J.R. Eds, (1998). *Test interpretation and diversity*. Washington DC: American Psychological Association.

Sebranek, P., Meyer, V., and Kemper, D. (1996). *Writers inc.: a handbook for writing and learning*. Lexington, MA: D.C. Heath and Company.

Sample Items with Bias, Fairness, and/or Sensitivity Concerns

1. Franco Piccione cooked spaghetti for his family. When he placed 6 ounces of pasta into the water, the water temperature was 160° F. After 3 minutes the water temperature reached 212° F. What information is **not** needed to find the mean (average) rate that the water temperature changed?

- A. 6 ounces of pasta
- B. 160° water temperature
- C. 3 minutes
- D. 212° water temperature

Type of Bias: _____

2. For a community service project, Amanda's class spent 2 hours at a retirement home. They spent $\frac{1}{2}$ of the time doing jigsaw puzzles, $\frac{1}{4}$ of the time reading, and the rest of the time watching television. How long did they spend watching television?

- A. 30 minutes
- B. 45 minutes
- C. 60 minutes
- D. 90 minutes

Type of Bias: _____

3. Which of the following items did Kathleen buy at the fair?

- A. a pretzel
- B. a snowball
- C. a slice of pizza
- D. a piece of pie

Type of Bias: _____

Samples of Items with Bias, Fairness, and/or Sensitivity Concerns (continued)

4. On July 1 the price of a share of Jolter Corporation stock was \$123.38. On January 1, the price of a share was \$97.41. What is the percent of decrease in the price of a share of Jolter Corporation stock between July 1 and January 1? (Round to the nearest hundredth.)

- A. 1.27%
- B. 21.05%
- C. 25.97%
- D. 26.67%

Type of Bias: _____

5. What is the main idea of the article?

- A. Doctors work long hours and neglect their wives and children.
- B. Doctors deal with many pressures in modern American society.
- C. Doctors pay a large amount of money to attend medical school.
- D. Doctors leave the profession more now than ten years ago.

Type of Bias: _____

Samples of Items with Bias, Fairness, and/or Sensitivity Concerns (continued)

6. What is the main conflict in the story?

- A. man versus man
- B. man versus nature
- C. man versus society
- D. man versus self

Type of Bias: _____

7. What did Eduardo learn from the visit with his grandfather?

- A. Age does not affect one's personality.
- B. Older people need help with everyday tasks.
- C. Age lessens one's appreciation for life.
- D. Older people often have special medical needs.

Type of Bias: _____

8. Compare Ken's experience with playing golf with a time when you played golf. Support your comparison with details from the story.

Type of Bias: _____

9. What does Kim enjoy most about summer?

- A. celebrating her birthday
- B. swimming at the lake
- C. playing softball at the park
- D. reading her favorite books

Type of Bias: _____

Samples of Items with Bias, Fairness, and/or Sensitivity Concerns (continued)

10. What could someone learn from reading the article?

- A. Mexican people often wear sombreros and eat tacos.
- B. Mexico has become a popular tourist destination.
- C. Mexican people are very friendly and helpful.
- D. Mexico produces many different kinds of fruit.

Type of Bias: _____

11. The Wampanoag people and the Pilgrims both lived in the same environment at the same time. Which is an example of a way the Indians used their environment **before** the Pilgrims arrived?

- A. dug wells
- B. grew corn, squash, and beans
- C. raised sheep for wool
- D. sawed trees into boards to build houses

Type of Bias: _____

12. According to the article, how is Marie different from the other children in her class?

- A. She likes to play the piano.
- B. She is a blind person.
- C. She is a tall person.
- D. She likes to work alone.

Type of Bias: _____

Samples of Items with Bias, Fairness, and/or Sensitivity Concerns (continued)

13. Samantha entered an ice-fishing contest. She drilled an 8-inch hole. What is the circumference of the hole Samantha drilled?

Use $\pi = 3.14$.

- A. 12.56 inches
- B. 25.12 inches
- C. 32 inches
- D. 50.24 inches

Type of Bias: _____

Use the table to answer question 14.

Favorite Sports	
Sport	Number of Members
Golf	13
Polo	9
Rugby	5
Sailing	21

14. Melissa conducted a survey at the Morningside Country Club. She asked members to name their two favorite sports. The table above shows the results. How many members are included in Melissa's survey?

- A. 12
- B. 21
- C. 24
- D. 48

Type of Bias: _____

Samples of Items with Bias, Fairness, and/or Sensitivity Concerns (continued)

15. In the story, Charlie says to Mia, “You can kill two birds with one stone.” What does this phrase mean? Use details from the story to support your answer.

Type of Bias: _____

16. Mrs. Sanders ordered new windows for her house. The salesman told her that each window would be made from 4 sections of glass. Which expression represents the number of sections of glass necessary to make W windows for Mrs. Sanders’ house?

A. $w + 4$

B. $w - 4$

C. $w \div 4$

D. $w \bullet 4$

Type of Bias: _____

Use the table to answer question 17.

Science Test Scores	
Student	Test Score
John	93
Susan	61
Tyson	96
Tao	93
Jessica	55
Takisha	70

17. The science test scores of six students in a lab group in Mr. Parker’s class are shown in the table above. What is the mean (average) score of the test scores in the lab group?

A. 41

B. 77.5

C. 80.5

D. 93

Type of Bias: _____

Samples of Items with Bias, Fairness, and/or Sensitivity Concerns (continued)

18. Animals are adapted to survive in their specific environments. Which breed of livestock would be best suited for meat production in the grassland prairies of the Great Plains?

- A. angora goat
- B. buffalo
- C. Hereford cattle
- D. Suffolk sheep

Type of Bias: _____

19. Scientists studying the fossil record have observed gradual changes in the structural morphology of numerous organisms that occurred over millions of years. These changes are most likely the result of

- A. accidents from cloning experiments.
- B. adaptive responses to environmental change.
- C. God's little mistakes leading to his creation of mankind.
- D. hallucinations of an occupant of the H.M.S. Beetle.

Type of Bias: _____

OE Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
----------	--

Prompt/Stem

Rubric Template

Item Writer:

Grade:

Alignment:

Scoring Guide:

Score	Description
2	This response demonstrates a <i>thorough</i> understanding of
1	This response demonstrates a <i>partial</i> understanding of
0	The response provides <i>insufficient</i> evidence to demonstrate any understanding of the concept being tested.
Non-scorables	B – Blank, entirely erased or written refusal to respond F – Foreign Language K – Off-task U – Unreadable

Responses that will receive credit:

OE Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
----------	--

Prompt/Stem

Rubric Template

Item Writer:

Grade:

Alignment:

Scoring Guide:

Score	Description
2	This response demonstrates a <i>thorough</i> understanding of
1	This response demonstrates a <i>partial</i> understanding of
0	The response provides <i>insufficient</i> evidence to demonstrate any understanding of the concept being tested.
Non-scorables	B – Blank, entirely erased or written refusal to respond F – Foreign Language K – Off-task U – Unreadable

Responses that will receive credit:

OE Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
----------	--

Prompt/Stem

Rubric Template

Item Writer:

Grade:

Alignment:

Scoring Guide:

Score	Description
2	This response demonstrates a <i>thorough</i> understanding of
1	This response demonstrates a <i>partial</i> understanding of
0	The response provides <i>insufficient</i> evidence to demonstrate any understanding of the concept being tested.
Non-scorables	B – Blank, entirely erased or written refusal to respond F – Foreign Language K – Off-task U – Unreadable

Responses that will receive credit:

OE Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
----------	--

Prompt/Stem

Rubric Template

Item Writer:

Grade:

Alignment:

Scoring Guide:

Score	Description
2	This response demonstrates a <i>thorough</i> understanding of
1	This response demonstrates a <i>partial</i> understanding of
0	The response provides <i>insufficient</i> evidence to demonstrate any understanding of the concept being tested.
Non-scorables	B – Blank, entirely erased or written refusal to respond F – Foreign Language K – Off-task U – Unreadable

Responses that will receive credit:

OE Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
----------	--

Prompt/Stem

Rubric Template

Item Writer:

Grade:

Alignment:

Scoring Guide:

Score	Description
2	This response demonstrates a <i>thorough</i> understanding of
1	This response demonstrates a <i>partial</i> understanding of
0	The response provides <i>insufficient</i> evidence to demonstrate any understanding of the concept being tested.
Non-scorables	B – Blank, entirely erased or written refusal to respond F – Foreign Language K – Off-task U – Unreadable

Responses that will receive credit:

MC Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
-----------------	--

Prompt/Stem

Answer Options	
Key:	
Option A.	
Rationale:	
Option B.	
Rationale:	
Option C.	
Rationale:	
Option D.	
Rationale:	

MC Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
-----------------	--

Prompt/Stem

Answer Options	
Key:	
Option A.	
Rationale:	
Option B.	
Rationale:	
Option C.	
Rationale:	
Option D.	
Rationale:	

MC Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
-----------------	--

Prompt/Stem

Answer Options	
Key:	
Option A.	
Rationale:	
Option B.	
Rationale:	
Option C.	
Rationale:	
Option D.	
Rationale:	

MC Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
-----------------	--

Prompt/Stem

Answer Options	
Key:	
Option A.	
Rationale:	
Option B.	
Rationale:	
Option C.	
Rationale:	
Option D.	
Rationale:	

MC Template

Item Writer Name:	Grade:
Standard:	Points:
DOK:	Estimated Difficulty:

Comment:	
-----------------	--

Prompt/Stem

Answer Options	
Key:	
Option A.	
Rationale:	
Option B.	
Rationale:	
Option C.	
Rationale:	
Option D.	
Rationale:	