

## **APPENDIX 3: ITEM WRITER ORIENTATION MANUAL**



# **Item Writer Orientation Manual**

## **Alaska Item Writer Orientation**

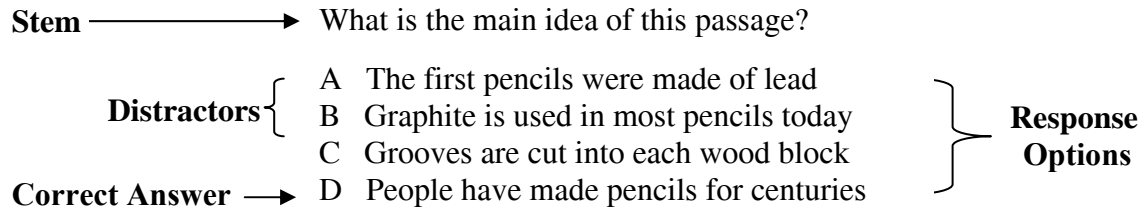
## **Table of Contents**

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
Bias and Sensitivity . . . . .	29
Topics to Avoid . . . . .	30
Special Considerations for Alaska items . . . . .	32
Glossary of Assessment Terms . . . . .	33

## Models of Multiple Choice and Constructed-Response Items

The Alaska Comprehensive System of Student Assessment uses 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.

<b>Rubric</b>	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

The Alaska Comprehensive System of Student Assessment includes two types of constructed response items. Selected Constructed Response (SCR) items are 2 point items, Extended Constructed Response (ECR) items are 4 point items or 6 point writing prompts.

## 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.

B. Better

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.

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, the 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, 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. 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  $1/2$
- B  $2/3$
- C  $3/4$
- D  $4/5^*$

B. Acceptable

Which has the greatest value?

- A  $3/4$
- B  $2/3$
- C  $4/5^*$
- D  $1/2$

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 Norman Webb's *Depth of Knowledge* 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

Who was George Washington’s vice-president?

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

B. Better

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 maker’s 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 Ivar 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\*

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

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

B. Better

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

## **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.

## Special Considerations for Alaska Items

1. **MANY ORDINARY FACTS IN THE “LOWER 48” ARE NOT TRUE IN ALASKA.** Examples-There are no snakes in Alaska! You cannot take a bus to its capital city, Juneau. In fact, you cannot even drive to Juneau. It is the only state capital accessible only by boat or plane. **Make sure you check facts for accuracy in Alaska.**

2. The following are special considerations unique to Alaska:

Air Conditioning

Skateboards

Skyscrapers

Snakes (there are no snakes in Alaska)

Mail Carriers (do not use; most Alaskans have a post office where they pick up their mail)

Fairs/carnivals (avoid with younger students; older students may be familiar with a state fair)

Garden Stores

Counties

School Auditoriums (don't use)

Blocks as in city blocks (avoid with 3<sup>rd</sup> grade)

Oak, Maple, Elm trees (Alaska has Alder and Birch)

Snowmobile (use snow machine)

Musk Oxen (use Musk Ox)

Fur Seal (use Seal)

Country, Countryside (use wilderness)

3. General Information:

The mountain is Mt. McKinley, the park is Denali

Costs are higher in Alaska (check prices)

Basketball is very common

Outside refers to out of Alaska; use outdoors.

Caution about passages or text regarding burial sites

Use a balance of urban and non-urban as in very, very, rural situations.

Latino names are rather uncommon.

Use sparingly-theaters (majority are limited to video rentals) Tennis, Football, Softball,

Museums, Parades, Parks (50% of students probably do not have a park in their area or live in a neighborhood.)

Check on use of city, town, village. There are a few towns (Kiana, Mountain Village, Kaktovik, Wasilla, Big Lake. There are more villages such as, Nulato, Alakanuk, Point Lay, Craig Wass, Gabriel, Jason, Sophie, Margie etc.

Use animals, birds, plants and trees common to Alaskans. Check the facts.

Common AK birds: swallow, robin, sea gull, raven, geese, swans, eagles, seagulls. Not really any cardinals, so they say.

## **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. Use conventional spelling of names. Use common names.

Common Alaska names: Anuska, Anecia, Natlaia, Tasha, Ivan, Evan, Wassillie, Nick, Josh, Herman, Annie, Anna, Peter, Travis, Noah, Constantine, Tory, Crystal, Helen, Sally, Sonny,

Common bush names: Cavelia, Gusty, Sonya, Dustin, Justin, Vanessa, Roberta, Tatiana, Bagriella, Cameron, Ester, Jimmy, Sanana

## **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

## **Alaska Content Standards**

Broad statements about what students should know and be able to do in core content areas.

## **Alaska Grade Level Expectations**

Specific expectations of what students should know and be able to do at specific grade levels.

## **Alaska Performance Standards**

Specific expectations of what students should know and be able to do at four levels: ages 5-7, 8-10, 11-14, and 15-18.

## **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.).

## **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. Constructed response items are of three types as follows:

- SCR–Short constructed response (2 points)
- ECR–Extended constructed response (4 points)
- Writing ECR–Writing prompt (6 points)

### **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.

### **Depth of Knowledge Levels**

The categorization of items according to the cognitive demand of the item. The four depth of knowledge levels are recall, basic application of skill/concept, strategic thinking, and extended thinking.

**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 determine what reasonably 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 median.

**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.

**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 (e.g., 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 25<sup>th</sup> percentile, the 26<sup>th</sup> to 50<sup>th</sup> percentile, the 51<sup>st</sup> to 75<sup>th</sup> percentile, and the 76<sup>th</sup> to 99<sup>th</sup> percentile.

**Quintile**

The breakdown of an aggregate of percentile rankings into five categories: the zero to 20<sup>th</sup> percentile; 21<sup>st</sup> to 40<sup>th</sup> 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 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.

**Universally Designed Assessments**

Assessments that allow the participation of the widest possible range of students. They are developed and designed to eliminate barriers that may disadvantage any particular group of students.

**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.