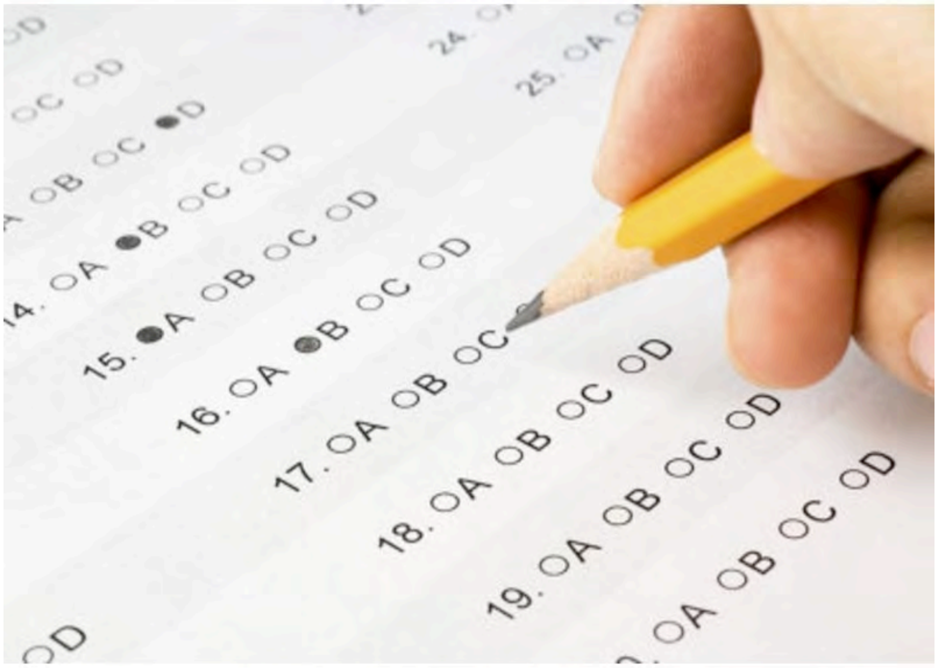


A Handbook for Improving the Validity of Multiple-Choice Science Test Items for English Language Learners

English Learners and Science Tests Research Project (ELAST)

Tracy Noble[1], Ann Rosebery, Rachel Kachchaf[2], Catherine Suarez
TERC

May, 2016

Table of Contents

[1] For additional information, contact Tracy Noble at Tracy_Noble@terc.edu
[2] Formerly of TERC, currently of Smarter Balanced Assessment Consortium.

Authors Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant # R305A110122. The opinions expressed herein are those of the authors and do not reflect the opinions of the funding agency. The authors would like to thank the many people who contributed to this work, including Beth Warren, Catherine O'Connor, Yang Wang, Catherine Bowler, Carol Lord, Richard Durán, Guillermo Solano-Flores, Joel Webb, Lori Likis, and Mary Rizzuto. We would also like to thank all of the Massachusetts students, teachers, and administrators who participated in and contributed to this work. This handbook is dedicated to them.

EXECUTIVE SUMMARY

This Handbook is based on research findings from The English Learners and Science Tests Research Project (ELAST).

Goals of Our Research

- To investigate features of multiple-choice items on a state science assessment that interfere with English Language Learners' (ELLs') abilities to demonstrate their science knowledge and skills.
- To improve standardized science test development by investigating methods to improve item writing.
- To contribute toward more fair and equitable testing in science, so that ELLs and their teachers, schools, and districts are not unfairly penalized as a result of test score gaps between ELLs and non-ELLs that result from differences in English language proficiency rather than science knowledge and skills.

Three Studies

We conducted three studies to investigate how the language on the Massachusetts (MA) Grade 5 Science and Technology/Engineering MCAS test (STE MCAS) might interfere with ELLs' abilities to demonstrate what they know on science tests:

1) Correlation Study: Investigated which of the linguistic features identified in previous research were highly correlated with ELL performance on Grade 5 Science MCAS multiple-choice items.
2) Interview Study: Interviewed 52 Grade 5 ELLs about test items with and without the linguistic features identified in the correlation study.
3) Test Administration Study: Modified released Science MCAS test items to remove problematic linguistic features and add helpful features. Administered original and linguistically modified test items to over 2000 Grade 5 students (ELLs and non-ELLs) in four MA districts.

Findings

Findings from these studies confirmed that the visual and linguistic content in MCAS science test items had an impact on ELL performance. The studies identified two features (one visual and one linguistic) that are helpful for ELLs and three linguistic features that cause difficulty for ELLs. These findings will be explored with recommendations and methods for how to improve multiple-choice items for ELL students.

About this Handbook

This Handbook presents research-based methods that aim to improve the development of multiple-choice science test items by reducing biases of item writing against ELLs. These methods are intended to help item writers develop items that primarily assess the scientific knowledge and skills of ELL students rather than their English language proficiency. Although the focus of our research was to reduce biases against ELLs, we have found these methods improve student achievement for all students. We urge science test developers, as well as those that review items for bias review and content, to use this handbook as a resource to help improve item development for all students.

I. Features that Affect the Performance of English Language Learners

A. Two Features that are Helpful to English Language Learners

1. Technical Vocabulary consists of individual words or compound terms that do not occur often in texts read by 5th grade students, but can nonetheless help ELLs perform better on the Grade 5 STE MCAS. Technical words have a primary scientific meaning or are used in a technical way in the *Massachusetts STE Curriculum Framework*. ELLs do better on items with these words because technical words have a clear scientific meaning, are generally taught in school science, and because they help students understand the correct context for answering the item. Example A below, Technical Vocabulary appears in blue.

Our definition of Technical Vocabulary for the purposes of our research did not include all scientific words, as it does not include science words that are used frequently in everyday speech and/or have a common meaning that is not based in a science discipline (e.g., *gas* and *sound*).

For this reason, test developers may wish to customize their definition of Technical Vocabulary along with their state Education Department partners, by identifying those words considered Technical Vocabulary for their test development purposes. Test developers may wish to in addition use the methods for identifying Technical Vocabulary that we describe below to identify any additional Technical words in test items.

Example A. (MA DOE, 2004): Electric Fan.

When an electric fan is running, **most** of the incoming electrical energy changes into which kind of energy?

A. heat energy
B. light energy
C. mechanical energy
D. sound energy

Identifying Technical Vocabulary. To determine whether a word is Technical, we recommend following procedures outlined by Wolf & Leon (2009)[3] and using the *Longman Dictionary of American English (2008)*, which lists definitions in order of frequency of use, with the most common definition given first. We recommend classifying a term as *Technical* if the first definition given in the Longman Dictionary is labeled with a science or mathematics discipline. A second step is to create a list of terms from the *Massachusetts STE Curriculum Framework* and classify them as "Technical" if the term has a science-specific definition in the *Framework* and students are likely to encounter it primarily in their science classes. Finally, we'd like to point out that some items may contain compound terms that qualify as "Technical Terms". We recommend defining a compound term as Technical if the meaning of the term is scientific and *not* predictable from the meaning of either of its

[3] Wolf, M. K. & Leon, S. (2009). An investigation of the language demands in content assessments for of English language learners. *Educational Assessment, 14*(3), 139-159.

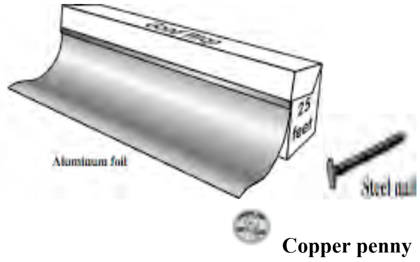
component, individual words. Examples of Technical Terms include *water cycle*, *food chain*, and *graduated cylinder*. The Technical words that appeared most often in multiple-choice items on the Grade 5 STE MCAS from 2004 to 2010 are: *climate, diagram, earth, electricity, erosion, igneous, life cycle, mineral, ocean, organism, plant, solid, stem, volcano,* and *luster.*

Technical vocabulary does not encompass all of the science vocabulary in test items. For example, there are many science words that are context-specific, that is, that have different meanings and uses in different contexts, such as the words *gas* and *sound.*

2. Visuals include any non-linguistic information found in test items such as pictures, tables, charts, and diagrams. Visuals can be found in either the question portion of an item (See Example B) or in the answer choices (See Example C). Visuals can be an important source of non-linguistic information for test-takers, particularly ELLs.

Example B. *(*MA DOE, 2004): Classifying Objects

The picture below shows three objects
that can be classified in the same group.



Which of the following statements is true for
all three of these objects?

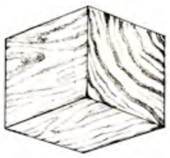
A. They are metals.
B. They rust rapidly.
C. They weigh the same.
D. They are the same color.

Example C. Flexible Object (ELAST modification of MA DOE item from 2005 that did not originally include visuals.)

Which of the following objects is probably the **most** flexible?

A. a ceramic dish

C. a short steel rod



B. a wooden block

D. A new rubber hose

Note: Our research shows that including visuals in multiple-choice science items has a significant effect on the performance of 5th grade ELL students. Directions for modifying items to include visuals can be found on pages 12-15.

B. Three Features that Cause Difficulty for ELLs

1. The Forced Comparison feature occurs in test items requiring students to compare all four answer choices and identify the one that expresses the correct extreme value of some variable.

- These items use extreme value terms like *best*, *most*, *most likely*, *greatest (*noted in red below in Example D). The specific meanings of these words depend upon the criteria for judging what is, for example, *best,* and the criteria the student should use to make this judgment are not always provided in the item.
- These items often use, in conjunction with the extreme value word, a verb, noun, or adjective that can have multiple meanings depending upon the context of use. Examples include *respond, help, explain, important, and effort* (verb used in conjunction with "most likely" is noted in green below in Example D).

Example D. (MA DOE, 2004): Earthworm

An earthworm was placed on top of a thick layer of moist topsoil in a pan. The pan was placed in a room with the lights on. How did the earthworm most likely respond to these conditions?

A. by burrowing under the soil
B. by crawling around in the pan
C. by staying where it was placed
D. by trying to crawl out of the pan

Many ELLs have difficulty defining extreme value terms such as *most likely,* and often use definitions related to other uses of the words "most" and "like" such as: *almost, most similar to, the one I like the most,* or *the one I don't like*. Many ELLs also use a definition such as "best answer" to define *most likely*, which generally does not help them to identify the correct answer. Without an understanding of the meaning of *most likely*, the student does not have criteria for choosing the "best answer".

In addition, the criteria that should be used to judge what is *best* or *most likely* are often implicit in Forced Comparison test items. Many students (both ELLs and non-ELLs) have difficulty inferring the criteria intended by the item writers. These criteria are often based upon assumptions of shared background knowledge that ELLs may not always have, such as knowledge of typical weather and climate in Massachusetts.

The noun, verb, or adjective that follows a term like *most likely*, *best*, or *greatest* may have multiple meanings or may be unclear in meaning. For the ELLs whom we interviewed, this was the case. In Example D, the verb "respond" was often interpreted by ELLs to mean: "answer when a person talks to you," "answer a question," or "give an answer to a test item" rather than "react to an environmental stimulus," the meaning presumably intended by the item writer.

Identifying Forced Comparison.

The first two bullets describe aspects of Forced Comparison items that are often, but not always present.

- The common question phrase, *Which of the following*.
- Set Y – What category of thing is being sought, as in: "Which of the following *environmental changes…"*

The second two bullets describe aspects of Forced Comparison items that are always present and are defining aspects of items with this feature.

- An End of Scale value, such as *best, most likely,* etc.
- A verb, noun, or adjective associated with the end of scale value, such as "Which of the following environmental changes most likely *caused a decrease…*"

Item writers should be aware that items that contain the Forced Comparison structure have been shown to disadvantage ELLs (Kachchaf et al, 2015[4]). A set of items on which to practice identifying the features discussed in this handbook appears on pages 19-22.

2. **The Reference Back** feature occurs in test items in which the question sentence refers back to information that appeared earlier in the item and that is needed to understand and solve the item. This feature can add difficulty for ELLs because it is not always clear what information is needed from earlier in the item, or even that one must look back in the item.

To determine if an item contains the Reference Back feature, try reading only the last sentence of the item and decide if you need to use information from the previous sentence(s) to correctly answer the question. If you cannot answer the question without that previous information, then the item contains the Reference Back feature.

Example E. (MA DOE, 2004): Earthworm

> An earthworm was placed on top of a thick layer of moist topsoil in a pan. The pan was placed in a room with the lights on. How did the earthworm most likely respond to these conditions?
>
> A. by burrowing under the soil
> B. by crawling around in the pan
> C. by staying where it was placed
> D. by trying to crawl out of the pan

In Example E, the term "these conditions" in the question refers back to the previous two sentences and the student must remember that the conditions for the earthworm include both the presence of a thick layer of moist topsoil in a pan (from the first sentence) as well as the placement of the pan in a room with the lights on (the second sentence).

[4] Kachchaf, R. R., Noble, T., Rosebery, A., Warren, B., O'Connor, M. C., and Wang, Y. (2015). A Closer Look at Linguistic Complexity: Pinpointing Individual Linguistic Features of Science Multiple-choice Items Associated with English Language Learner Performance. *Manuscript in preparation*.

Example F. (MA DOE, 2008): Marsh Willow Herb

> The marsh willow herb is a plant native
> to the northeastern United States. It grows
> best in damp habitats. Which of the
> following environmental changes would
> most likely cause a decrease in the marsh
> willow herb population in an area?
>
> A. a rainstorm lasting several weeks
> B. a drought lasting twelve months
> C. unusually low temperatures during the
> month of July
> D. unusually high temperatures during the
> month of January

In Example F, a student must use the information about what a marsh willow herb is (from the first sentence) and how it grows best (from the second sentence) in order to answer the question posed in the last sentence, even though there is no explicit reference to prior sentences.

Identifying Reference Back. The Reference Back feature occurs in test items in which the question sentence requires the test-taker to refer back to information from previous sentences in the question that is needed to understand and solve the item. Given that a student understands the vocabulary of the item, he/she must integrate what the question is asking with information provided prior to the question in order to answer the question. There may be an explicit anaphoric reference, which means the question sentence uses a pronoun to refer back to a noun in the sentence. In other cases, the question sentence may refer back without the use of a pronoun.

To determine whether an item contains the Reference Back feature, we recommend that two coders with experience in science education and assessment independently read only the question sentence of the item and any visuals contained in the item. If the readers can confidently answer the item without needing the rest of the text of the item, the item does not contain the Reference Back feature. If the readers can NOT confidently answer the question and need information provided in a previous sentence or sentences to answer the item, the item has the Reference Back feature. Item writers should be aware that items that contain the Reference Back feature have been shown to disadvantage ELLs (see Kachchaf et al, 2015 as noted above). A set of items on which to practice identifying the features discussed in this handbook appears on pages 19-22.

3. Low Frequency Non-Technical Vocabulary consists of words that do not appear routinely in 5th grade texts. These words are not technical and include many words that are not thought of as “scientific”, and therefore may not be taught in science class, such as *approximate, classified*, *incomplete,* and *toaster*.

In our coding scheme, created for research purposes, this category also includes many science terms that have context-specific meanings, such as *consumer*, or are frequently used outside of science, such as *conduct*, *crystal*, and *drought*. It is important to note that while such science words may be part of the elementary school science standards, science words that have multiple meanings and uses outside of science may be particularly challenging for ELLs. For words that may be part of elementary school science standards but also have multiple uses and meanings outside of school, such as those listed above, it is important to verify that the context of the test item containing such words helps students to identify the correct meaning and use of the word.

Test developers who have worked with their state's Department of Education to identify what is considered Technical Vocabulary should use that list of words to customize their list of Low Frequency Non-Technical Vocabulary.

In Example G below, Low Frequency Non-Technical Vocabulary is highlighted in red.

Example G. (MA DOE, 2008): Marsh Willow Herb.

The marsh willow herb is a plant native to the northeastern United States. It grows best in damp habitats. Which of the following environmental changes would most likely cause a decrease in the marsh willow herb population in an area?

A. a rainstorm lasting several weeks
B. a drought lasting twelve months
C. unusually low temperatures during the month of July
D. unusually high temperatures during the month of January

Identifying Low Frequency Non-Technical Vocabulary. To identify Low Frequency Non-Technical Vocabulary, we used the *Educator's Word Frequency Guide* (Zeno et al., 1995[5]), a word frequency database used widely. While this database is not as current as we would have liked, it was the best resource available at the time of our study. This database compiled many texts at a number of grade levels and counted the number of times a given word appeared in these texts. At the Grade 5 level, a word was classified as infrequent if it appeared less than 10 times per million in Grade 5 texts. It is important to point out that the database considers different forms of the same word (e.g., cat and cats) as different words. Following Bauer and Nation (1993)[6], who argue that learners familiar with one form of a word like *cat* are likely to be familiar with other forms, e.g., *cats*, we recommend classifying regular word forms as the same word if they have the endings provided in the table below.

Bauer and Nation's Level 2 Classification of Word Forms

Category	Change in spelling	Example	Is NOT
Plurals	+ *s* or + *es*	book – books	foot - feet
Past tense	+*ed*	fine-fined	eat - ate
Gerund	+ *ing*	Doing	doings (noun)
Third person singular	+*s*	Eats	have - has
Comparative	+*er*	high – higher	consume - consumer
Superlative	+*est*	high – highest	
Possessive	+'*s*	bear's or bears'	

We understand that it would be impractical to look up every word that might be considered Low Frequency Non-Technical and apply this method to each. However, in our experience, having an understanding of the types of words that might be classified in this way can go a long way in editing an item that you feel might have these words. The more that you go through these items with a critical eye towards words that might be challenging for ELLs, the easier it will be to eliminate non-technical words that might be problematic to ELL students (because the words are unfamiliar, have multiple non-scientific meanings, etc.). These words can then be replaced with words that are likely to be more familiar to ELLs without changing the science being tested by the item. A set of items on which to practice identifying the features discussed in this handbook appears on pages 19-22.

[5] Zeno, S., Ivens, S. H., Millard, R. T., & Rothkopf, E. Z. (1995). *The educator's word frequency guide*. Brewster, NJ: Touchstone Applied Science Associates.

[6] Bauer, L. & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

II. Methods for Modifying Items to Include Visuals

In this section, we explain three methods for modifying items to include visuals that improve the performance of ELLs:

Method A. Adding a visual to the stem of an item, page 12
Method B. Adding visuals to the answer choices of an item, page 13
Method C. Maximizing visuals already present in an item, page 15

Visual information includes a wide array of pictures, tables, charts, diagrams, etc. The purpose of adding visual information to an item is to increase the non-linguistic information available to English learners, and thus provide parallel sources for making meaning. Thus, most items that do not contain these elements can benefit from the addition of visual information.

One exception to this rule is items that involve a process or an event that changes over time. Our research shows that processes and events across time are difficult to illustrate with visuals, and may even make such items more confusing for students because an illustration typically depicts a single moment in time. Thus, we do not recommend adding visuals that attempt to depict processes or events that unfold across time.

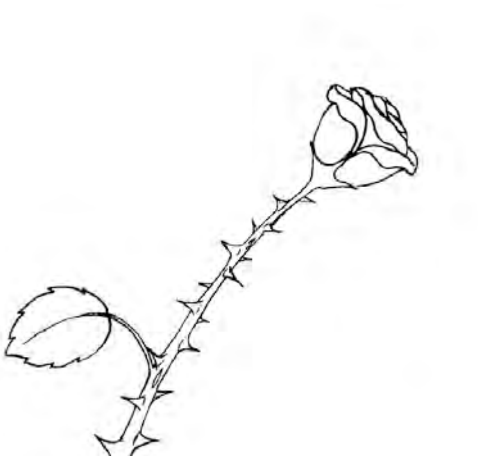
When modifying a given item it is important to maintain the relative difficulty level of the science that is being tested by that item (i.e., don't make the science content tested by the item "easier"). To ensure that the tested science content stays the same, we recommend the addition of visuals that do *not* give away the answer to the item or aspects of the science knowledge targeted by the item. Note that providing images may appear to make an item "easier" because the task of linguistic comprehension is easier, but may still maintain the same tested science content. To check to see whether the difficulty level of the science in an item has been maintained during the modification process, please refer to *Section III. Coding the Science Content of Items* on page 17. We recommend checking the science content knowledge and skills targeted by a test item before and after any modifications are performed to ensure that level of difficulty of the science task has not changed after modification.

Method A. Adding a Visual to the Stem

1. Select an item with no visual component in the stem (introductory information in the item).
2. Identify a noun in the item that is *not* included in the targeted science knowledge from the standard associated with the item.
3. Create a visual to include only the components necessary to illustrate the noun, avoiding extraneous details.

An Example. To answer the item below, students must know the word *thorns.* This word is not included in the science knowledge targeted by the item, as listed in the standard, which does not mention thorns. According to our research, the word thorn is often learned by students outside of school, and therefore may be unfamiliar to many ELLs. The modified version added a picture of a rose with thorns to illustrate this word. The modified version also added an introductory sentence to refer students to the illustration. Thus, by adding an illustration of a rose with thorns, the modified version provides ELL students with important non-linguistic information (i.e., what thorns are). However, it does not affect the difficulty level of the science being tested because to answer correctly, students must know what the *purpose* of thorns on a plant is.

Example H, aligned with Life Science, Learning Standard 2. Identify the structures in plants (leaves, roots, flowers, stem, bark, wood) that are responsible for food production, support, water transport, reproduction, growth, and protection.

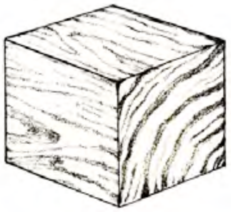
Original	**Modification: Visual added to the stem**
The purpose of thorns on a plant is most likely to A. help the plant to get some moisture. B. anchor the plant in the ground. C. protect the plant from harm. D. support the stems and branches.	The picture below shows a plant with thorns.  The purpose of thorns on a plant is most likely to A. help the plant to get some moisture. B. anchor the plant in the ground. C. protect the plant from harm. D. support the stems and branches.

Method B. Adding Visuals to the Answer Choices

1. Select an item with no visual component in the answer choices (includes pictures, graphs, and tables).
2. Identify a noun in the item that is *not* included in the targeted science knowledge from the standard associated with the item.
3. Create a visual to include only the components necessary to illustrate the noun in each answer choice, avoiding extraneous details.

To answer the item below, students must know all of the words in the answer choices (e.g., *ceramic dish, wooden block*). These words are not part of the science knowledge targeted by the standard and therefore modifying the item by adding visuals to the answer choices will not change the difficulty level of the item.

Example I, aligned with Physical Sciences, Learning Standard 1: Differentiate between properties of objects (e.g., size, shape, weight) and properties of materials (e.g., color, texture, hardness).

Original	**Visual modification: Adding visuals to answer choices**
Which of the following objects is probably the most flexible? A. a ceramic dish B. a wooden block C. a short steel rod D. a new rubber hose	Which of the following objects is probably the most flexible? A. a ceramic dish C. a short steel rod  B. a wooden block D. A new rubber hose

Method C. Maximizing Visuals already Present in an Item

1. Select an item that contains visuals in the forms of charts or tables.
2. Determine if the information in the table is not part of the science knowledge assessed.
3. Create visuals that include only the components necessary to illustrate the information in the item, avoiding extraneous details.

To answer the item below, students must know all of the words in the lists. These words are not part of the science knowledge targeted by the standard and therefore modifying the item by adding visuals to the stem will not change the difficulty level of the item.

Example J, aligned with Life Science, Learning Standard 1: Classify plants and animals according to the physical characteristics that they share.

Original

The lists below show the organisms that Tamara sorted into two groups based on one physical characteristic.

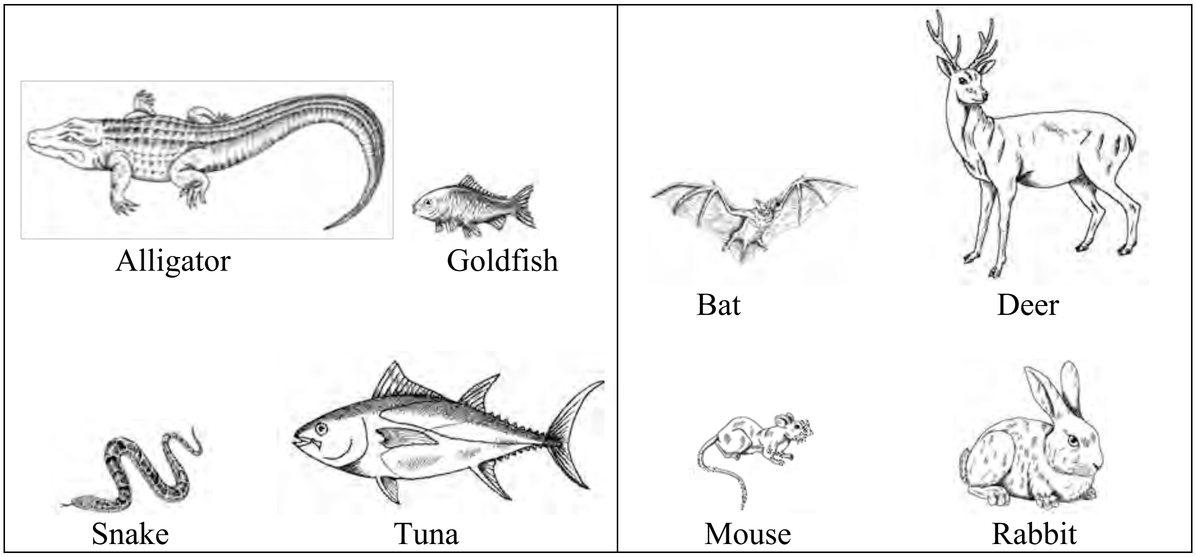
Group 1	Group 2
alligator	bat
goldfish	deer
snake	mouse
tuna	rabbit

Which of the following physical characteristics did Tamara most likely use to sort the organisms into the two groups?

A. number of legs
B. size of the body
C. shape of the feet
D. type of body covering

Modified

The pictures below show the organisms that Tamara sorted into two groups based on one physical characteristic.



Which of the following physical characteristics did Tamara most likely use to sort the organisms into the two groups?

A. number of legs
B. size of the body
C. shape of the feet
D. type of body covering

III. Coding the Science Content of Items

This protocol was designed to code the science content on the Grade 5 STE MCAS. In the ELAST project, two science coders who were district science coordinators with extensive experience of 5th grade science curricula in Massachusetts answered the following list of questions for each original test item and for each modified test item. The ELAST team used the answers provided by the science coders to judge whether the modified items tested the same science content as the original items and made changes to the modifications when the science content had been inadvertently changed. This protocol has not yet been tested with science test items from other state or national tests and may need modifications in order to encompass the range of science content and skills assessed be a specific test.

List of Questions

Read the test item and select the correct answer: ___ (Check answer key if needed.)

1) What is the science knowledge you would expect most 5th grade students to use to choose the correct answer?

2) Is the science content of the item likely to be taught in grades 3-5 in your experience? (Y/N) Explain if necessary.

3) Is there anything about the visual images in the item that may be difficult for 5th grade students to interpret even if they know the science of the item? If yes, please describe.

[This doesn't have to be answered for every item with a visual display. Only ones that you believe are likely to cause problems for students because they are non-standard or complex, or for other reasons.]

4) Please give the **Science Task code(s)** for the item, from the list of task codes (see Appendix, *Task Codes List for Coding Science Content* for codes developed and used in ELAST research), and any explanation you feel is needed.

Science Standard for the item: Find the standard keyed to the test item in the published set of released items, usually at the back.

5) Please state whether the science content described by the science standard is Necessary to answer the item correctly (Y/N). Explain if necessary.

6) Please state whether the science content described by the science standard is Sufficient to answer the item (Y/N). Explain if necessary.

7) Additional Comments:

IV. Features Checklist

The following checklist can be used to determine if an item contains the features outlined in this handbook. If an item contains Forced Comparison, Reference Back, or Low Frequency Non-Technical Vocabulary, we suggest that it be modified if possible. If the item can be enhanced by the addition of Technical Vocabulary or Visuals, we suggest that these features be added. This checklist provides a general overview of the features that may be present in any given item and therefore can serve as a reminder of how those features might affect ELL performance.

Features that are Challenging for ELLs	Yes	No
Does the item contain Forced Comparison feature? (e.g. does it contain an extreme value term like *best, most, most likely, least likely, greatest*, etc.?)		
Can the item be rewritten to eliminate this feature?		
Does the item contain the Reference Back feature? (Reminder: Read ONLY the last sentence of the item, and if you cannot answer the question without information in the previous sentences then the item contains a Reference Back feature.)		
Is it possible to rewrite the question to include additional information to clarify an vague reference back term such as "situation" or …		
Does the item contain Low Frequency Non-Technical (LFNT) Words? Or words that a Grade 5 student would not be exposed to in an elementary science class?		
Features that Are Helpful for ELLs		
Does the item contain Technical Vocabulary? (words, based on the state standards, that would be expected to be taught in an elementary science class)		
Does the item contain visuals (including pictures, tables, charts, and diagrams) in either the stem or answer choices?		
Can the item be modified to include visuals for words that Grade 5 students may not be familiar with or are not expected to be taught in an elementary science class?		

V. Additional MCAS Science Test Items for Practice

These items[7] contain one or more of the following features: Forced Comparison, Reference Back, Low Frequency Non Technical Words, and Technical Words. Use the methods outlined in Section II to identify the presence of these features. (An answer key to identifying the presence of these features appears on page 21.)

2004

14) Ricardo has an igneous rock in his rock collection. Where did this rock **most likely** form?

A. in a volcano
B. on a forest floor
C. on a coral reef
D. at the bottom of a river

27) Which of the following organisms have the **greatest** effect on the ecosystem because of the changes they make to the environment?

A. bees building a hive in a hollow tree
B. wasps building a nest in a leafy bush
C. beavers building a dam across a stream
D. fish digging a burrow on a river bottom

[7] All the items are released items retrieved from Massachusetts Department of Elementary and Secondary Education web site (http://www.doe.mass.edu/mcas/testitems.html).

2005

22) What happens to the path of a light ray as it passes from air into water at an angle?

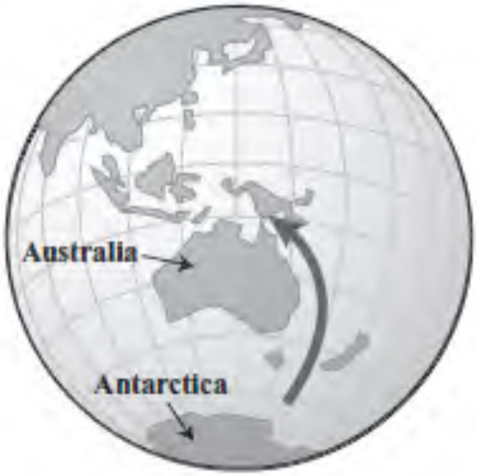
A. Its path widens.
B. Its path bends.
C. Its path becomes shorter.
D. Its path continues in a straight line.

27) Clouds and fog are made up of

A. water.
B. heat.
C. light.
D. helium.

2007

7) Each year, humpback whales migrate from the coast of Antarctica to the north coast of Australia. The map below shows the whales' migration route.



Which of the following are the whales most likely responding to when they begin to migrate?

A. the force of gravity
B. a shift in ocean waves
C. a change in water temperature
D. the approach of stormy weather

12) Sandra puts some pill bugs into an open box. She covers half the box with a piece of cardboard. She then places the box outside on a summer day, and all the pill bugs move under the cardboard.

The pill bugs are **most likely** responding to which of the following?

A. air pressure
B. bright light
C. wind
D. fog

2010

1) Francis plugged a toaster into an electric outlet. He put a piece of bread in the toaster and turned the toaster on. While the toaster was on, it changed the electrical energy from the outlet into other forms of energy.
Which form of energy toasted the bread?

A. chemical
B. heat
C. magnetic
D. sound

12) Owen tested a physical property of a mineral. He rubbed a mineral sample on a piece of white tile. The mineral left a red mark on the tile.

Which of the following physical properties of the mineral was Owen **most likely** testing?

A. cleavage
B. hardness
C. luster
D. streak

15) Corals are small marine organisms that live in groups and make hard outer skeletons to protect their bodies. Over time, these outer skeletons can build up to make large coral reefs.

Which of the following statements **best** describes one way the formation of a coral reef changes the ocean ecosystem?

A. It makes the ocean water saltier.
B. It removes sand from the ocean floor.
C. It causes ocean waves to become stronger.
D. It creates a habitat for some ocean animals.

Answer Key to Practice Items

The following items contain the Forced Comparison feature only: 2004 #14, 2004 #27, 2007 #12, and 2010 #12

The following features contain both Forced Comparison AND Reference Back: 2007 # 7, 2009 #4, 2010 #15,

Low Frequency Non Technical Vocabulary appear in red.

Technical Vocabulary appears in blue.

2004

14) Ricardo has an igneous rock in his rock collection. Where did this rock **most likely** form?

A. in a volcano
B. on a forest floor
C. on a coral reef
D. at the bottom of a river

27) Which of the following organisms have the **greatest** effect on the ecosystem because of the changes they make to the environment?

A. bees building a hive in a hollow tree
B. wasps building a nest in a leafy bush
C. beavers building a dam across a stream
D. fish digging a burrow on a river bottom

2005

22) What happens to the path of a light ray as it passes from air into water at an angle?

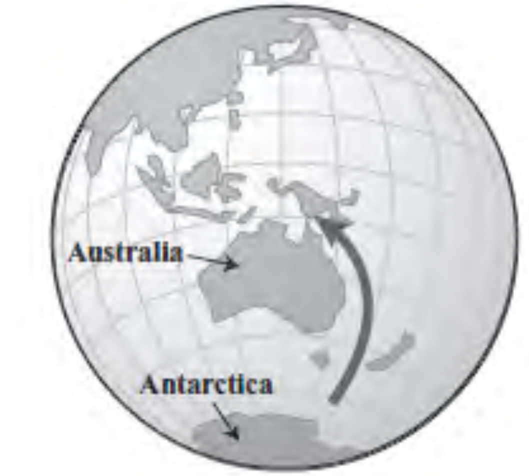
A. Its path widens.
B. Its path bends.
C. Its path becomes shorter.
D. Its path continues in a straight line.

27) Clouds and fog are made up of

A. water.
B. heat.
C. light.
D. helium.

2007

7) Each year, humpback whales migrate from the coast of Antarctica to the north coast of Australia. The map below shows the whales' migration route.



Which of the following are the whales most likely responding to when they begin to migrate?

A. the force of gravity
B. a shift in ocean waves
C. a change in water temperature
D. the approach of stormy weather

12) Sandra puts some pill bugs into an open box. She covers half the box with a piece of cardboard. She then places the box outside on a summer day, and all the pill bugs move under the cardboard.

The pill bugs are **most likely** responding to which of the following?

A. air pressure
B. bright light
C. wind
D. fog

2010

1) Francis plugged a toaster into an electric outlet. He put a piece of bread in the toaster and turned the toaster on. While the toaster was on, it changed the electrical energy from the outlet into other forms of energy.
Which form of energy toasted the bread?

A. chemical
B. heat
C. magnetic
D. sound

12) Owen tested a physical property of a mineral. He rubbed a mineral sample on a piece of white tile. The mineral left a red mark on the tile.

Which of the following physical properties of the mineral was Owen **most likely** testing?

A. cleavage
B. hardness
C. luster
D. streak

15) Corals are small marine organisms that live in groups and make hard outer skeletons to protect their bodies. Over time, these outer skeletons can build up to make large coral reefs.

Which of the following statements **best** describes one way the formation of a coral reef changes the ocean ecosystem?

A. It makes the ocean water saltier.
B. It removes sand from the ocean floor.
C. It causes ocean waves to become stronger.
D. It creates a habitat for some ocean animals.

V. Summary of Results for Three Studies

Study 1, Correlation Study: In Study 1, we investigated the relationship between the presence of the five linguistic and visual features defined in this Guide and the performance of English language learners (ELLs) on 162 multiple-choice items on Grade 5 STE MCAS. We used a statistic called Differential Item Functioning (DIF) to identify the items with DIF favoring non-ELLs over ELLs, that is, items on which ELLs' scores compared to non-ELLs' scores were lower than expected given their scores on the other items on the test. The presence of individual features was correlated with item DIF values, as shown in Table 1. We found two features significantly correlated with *lower* levels of DIF favoring non-ELLs over ELLs (negative correlation coefficients in Table 1): Technical science terms and Visuals. In other words, ELLs did better than expected on items with these features, when compared to non-ELLs. We also found three features that were correlated at a statistically significant level with *higher* levels of DIF favoring non-ELLs over ELLs (positive correlation coefficients in Table 1): Low Frequency Non-Technical vocabulary, Reference Back, and Forced Comparison. In these cases, ELLs did worse than expected on items with these features, when compared to non-ELLs.

In these three studies, we used the Massachusetts's classification of students as Limited English Proficient (LEP), based on a home language survey and English proficiency assessment, to classify students as ELLs. In addition, we analyzed the experience of students who had been moved out of LEP status within the previous two years, designated as Formerly LEP, or FLEP, in Massachusetts, because many FLEP students face challenges when answering MCAS science items in English that are similar to the challenges faced by LEP students.

Table 1. Effects of Linguistic Features on the Performance of English Language Learners on Grade 5 Science MCAS Multiple Choice Items for Correlation Study

Correlation of Linguistic Feature and Item DIF Value

Feature	LEP	FLEP
Technical Science Terms	-.206**	-.106
Visual	-.276**	-.197*
Low Frequency non-Technical Vocabulary	.155*	.146
Forced Comparison	.194*	.192*
Reference Back	.192*	.101

Notes: * = p<.05, ** p <.01.

Study 2, Interview Study: In Study 2, we interviewed 52 Grade 5 ELLs about 32 test items with and without the five linguistic features identified in the correlation study. Forty-one of these students were classified as LEP by their schools, and ten[8] of these students were classified as FLEP by their schools. Each student was asked to answer six test items and then immediately interviewed on at least four of those items by a bilingual interviewer who spoke the same first language as the student. Students were asked a range of questions based on the items, including: what answer choice they picked and why, what they thought the meanings of the target vocabulary/phrases were (e.g., Low Frequency non-Technical (LFNT) words, phrases related to the Forced Comparison feature such as "most likely," technical words, etc.), if and how they

[8] We were not able to obtain classification data for one student.

used the visual information, where they had learned about the subject matter (e.g., this year in school, in an earlier grade, at home) and what they had learned about it, and how they would change the item if they could rewrite it to make it more understandable.

Responses to the questions related to LFNT words and Forced Comparison items were coded as either correct, partially correct, or incorrect based on whether the student's interpretations of the language of these items matched the interpretations intended by the item writers. Table 2 shows the percentage of LEP and FLEP students who did not provide the intended interpretations for these words and phrases. The results demonstrate that both LEP and FLEP students had considerable difficulty with the language related to Forced Comparison items. And, as expected, LEP students provided the intended interpretations less often than FLEP students for both LFNT vocabulary words and the language related to Forced Comparison items.

Table 2. Effects of Linguistic Features on the Performance of English Language Learners on Grade 5 STE MCAS Multiple Choice Items for Interview Study

	Interview Results *(% of cases in which students did not give intended definition)*	
Feature	LEP	FLEP
Low Frequency non-Technical Vocabulary	39%	21%
Forced Comparison	78%	57%

Study 3, Test Administration Study: In Study 3, we modified 22 released MCAS test items to add Visuals (identified in Study 1 as a helpful feature) and to remove three problematic linguistic features: Forced Comparison, Reference Back, and Low Frequency Non-Technical vocabulary (as identified in Studies 1 and 2). We administered test forms consisting of both original and modified test items to over 2000 Grade 5 students (LEP, FLEP, and non-LEP) in four MA districts. Tests were analyzed as to whether or not students' performance improved on modified items.

For some items, pairs of features were modified together. We modified the Forced Comparison and Low Frequency Non-Technical vocabulary features together as one type of modification, and we modified the Forced Comparison and Reference Back features together as another type of modification. No additional modifications were made when a Visual was added to a test item. Table 3 shows the change in scores caused by each of these three types of modifications. We found that the addition of a Visual led to a statistically significant increase in LEP students' scores on test items, but that removing the problematic features in pairs did not lead to a statistically significant change in scores. The addition of a Visual also led to a statistically significant increase in the scores of non-LEP students who scored below proficient on the MCAS English Language Arts (ELA) test (a group that includes some FLEP students), while removing the problematic features did not. We conclude that while other evidence shows that the LFNT, Forced Comparison, and Reference Back features were problematic for LEP and FLEP students, modifying these features in pairs was insufficient to improve students' performance on these test items.

Table 3. Difference in Scores for LEP and non-LEP Below Proficient students on Original and Modified Items from the Grade 5 STE MCAS by Modification Type for Modification Study

Modification Type	*LEP*	*Non-LEP Below Proficient on MCAS ELA*
Visual	+3.3%*	+6.7%*
LFNT & FC	-0.9%	+2.5%
FC & RB	+0.0%	+0.7%

Note: * = p<.05

VII. For Those Who Would Like to Learn More

Please visit https://external-wiki.terc.edu/display/CKC/Publications for available conference papers and research reports as well as links to journal publications, and contact Tracy Noble at Tracy_Noble@terc.edu for further information.

Kachchaf, R. R., Noble, T., Rosebery, A., Warren, B., O'Connor, M. C., & Wang, Y. (2016). A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner performance. *Bilingual Research Journal*. *Manuscript in press*.

Noble, T., Bowler, K., Kachchaf, R., Sireci, S., Rosebery, A., & Wang, Y. (April, 2016). *Addressing the linguistic challenges of assessing English Learners: A state and research organization partnership.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Noble, T., Kachchaf, R. R., & Rosebery, A. (2016). Assessment and English language learners: Synthesizing research on linguistic features and construct- irrelevant variance. *Manuscript in revision*.

Kachchaf, R. R., Noble, T., Rosebery, A., Wang, Y., Warren, B., & O'Connor, M. C. (April, 2014). *The impact of discourse features of science test items on ELL performance*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia.

Noble, T., Kachchaf, R. R., Rosebery, A., Warren, B., O'Connor, M. C., & Wang, Y. (April, 2014). *Do linguistic features of science test items prevent English Language Learners from demonstrating their knowledge?* Paper presented at the annual meeting of the National Association of Research on Science Teaching, Pittsburgh.

Noble, T., Rosebery, A., Kachchaf, R., & Suarez, C. (2015). Lessons Learned and Implications for Practice from the English Learners and Science Tests Project. Unpublished manuscript. TERC, Cambridge, MA.

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*(4), 248-260. doi: 10.1080/08957347.2014.944309

Noble, T., Sireci, S., Wells, C. Kachchaf, R., Rosebery, A., & Wang, Y. (April 2016). *Targeted linguistic modifications of science items for English Learners.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching, 49*(6), 778-803. doi: 10.1002/tea.21026. **(Featured in** National Science Teachers' Association's 'Research worth reading: Summer 2013', *Journal of College Science Teaching*, *42*(6).

Appendix: Task Codes List for Coding Science Content

This list of task codes was developed from the Massachusetts Science and Technology/ Engineering Curriculum Framework (MA DoE, May 2001). We used it to identify the science tasks required to answer 62 Grade 5 multiple-choice Science MCAS items from the Earth and Space Science, Life Science, and Physical Sciences standards that were released in the years 2004-2010 and that were identified as potential items for modification.

This list of task codes may not be applicable to a larger set of Science MCAS items or to other test items developed according to new standards. However, it may serve as a useful starting point for the development of a taxonomy of science task codes.

1) Classification
    a. Identify the physical characteristics, properties, places of origin, method of formation, constituents, function, or other defining features that are used to classify animals, plants, rocks, objects, behaviors, traits, climates, substances, etc. in a specific way.
    b. Identify the animals, plants, rocks, objects, behaviors, traits, climates, substances, etc. that would (or would not) fall under a given category name.
    c. Identify the category into which animals, plants, rocks, objects, behaviors, traits, climates, substances, etc. with specific places of origin, method of formation, constituents, function, or other defining features would fall.

2) Purpose or Function
    a. Physical or Earth Sciences: Choose the purpose or function of a human-made object or part of a human-made object.
    b. Life Sciences: Choose the structure that goes with a particular function.
    c. Life Sciences: Choose the function or benefit of a particular structure, ability, or behavior.
    d. Identify physical characteristics or behaviors of plants or animals that contribute to their abilities to survive, to survive specific situations, or to perform specific functions.
    e. Identify how a physical characteristic or a behavior of a plant, animal or object fulfills a purpose, that is, in what way does this physical characteristic or behavior allow the plant or animal or object to, for example, act in a certain way, or withstand certain environmental forces.

3) Cycles: Recall and make inferences about cycles.
    a. Identify the correct name of a part of a cycle, given the location (in the order of parts or in time) in the cycle.
    b. Identify the correct location of a part of a cycle whose name is given.
    c. Identify the correct order for a list of parts of a cycle.
    d. Identify a fact that is not listed in (a) about a cycle or part(s) of a cycle.

4) Procedures
    a. Given a procedure, choose its purpose.

b. Given a purpose, choose the correct procedure or experiment or a step in the procedure or experiment, to accomplish the goal.

5) Energy Transformation: Identify the input or output or both in a process of energy transformation.

6) Co-varying quantities: Identify how one of two co-varying quantities changes if the other changes. Physical Sciences: How pitch changes with string length, tightness, and similar quantities for other musical objects. Life Sciences: How height changes with age.

7) Cause and Effect: This category describes a range of test items that ask about one event that will lead to, would lead to, or has led to another event. A cause may be a single event or ongoing process, and may be one of a number of contributing causes of a given effect or conditions necessary for an effect to occur. An effect can be a direct or indirect outcome of the cause and can be one of multiple effects. Examples of causes, given this broad definition, can include: erosion due to water, drought, the evaporation of water, the heating of water, the cooling of water, the presence of a complete circuit (for a light bulb to light), etc. Examples of effects can include: the survival of an organism or group of organisms, the reduction in the population of an organism, the lighting of a light bulb, the appearance of sugar crystals in a dish.

   a. Identify a cause (i.e., a single event, an ongoing process, or one of a number of contributing causes of a given effect or conditions necessary for an effect to occur) of a given effect or effects.
   b. Identify the effect(s) of a given cause (can be a single event or ongoing process, or one of a number of contributing causes of a given effect or conditions necessary for an effect to occur).

8) Graph or Table: Necessitates the use of information from a graph or table-like representation. Does not include reading information from diagrams or pictures.

9) Recall fact: Recall one or more specific facts about a process or object that is not one of the types of facts listed in any of the above categories. Check to see if the item fits into any of the previous 9 categories before choosing this category. For example: the fact that the earth is the 3rd planet from the sun.

10) Other: Does not fit into any of the previous categories. Write a description of the item task.