Lessons Learned and Implications for Practice from the English Learners and Science Tests Project

A Guide for Teachers

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

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

Table of Contents

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 guide is dedicated to them.

EXECUTIVE SUMMARY
The English Learners and Science Tests Project

Goals

- To make sure that MCAS science tests are testing students' knowledge and skills in science and not acting as additional tests of English proficiency.
- To make science testing more fair and equitable, so that English Language Learners (ELLs) and their teachers, schools, and districts are not unfairly penalized as a result of test score gaps between ELLs and non-ELLs.

Three Studies

We conducted three studies to investigate how the language of the 5th grade Science and Technology/Engineering MCAS might interfere with ELLs' abilities to demonstrate what they know about science when answering multiple-choice items:

1) **Correlation Study**: Investigated which of the linguistic features identified in previous research were highly correlated with ELL performance on 5th grade multiple-choice MCAS science items.
2) **Interview Study**: Interviewed Grade 5 ELLs about the test items with and without the linguistic features identified in the correlation study.
3) **Test Administration Study**: Modified released 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.

Study Findings

Findings from these studies confirmed that the visual and linguistic content in MCAS science test items had an impact on ELLs' performance on these science test items. (A summary of the results can be found on pages 15-17.)

Specifically, the studies:

1) Identified 2 features of 5th grade multiple-choice MCAS science test items that are helpful for ELLs.
2) Identified 4 linguistic features of 5th grade multiple-choice MCAS science test items that cause difficulty for ELLs.

These features, along with released items that illustrate them, appear on pages 4-8. (Additional items containing these features appear on pages 13-14.) We created two tip sheets to help teachers of ELLs translate our findings into practice. One tip sheet focuses on ways to prepare your students to take standardized tests (see page 9). The other focuses on best practices for developing your own assessments for use with ELLs (see page 11).

Two Features that are Helpful to ELLs

1. Technical Words are individual words or compound terms that do not occur very often in texts read by 5th grade students, but can nonetheless help ELLs perform better on the 5th grade STE MCAS. Technical words have a primary scientific meaning or are used in a technical way in the Massachusetts STE Frameworks. In Example A below, Technical Words appear in blue.

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, luster.*

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

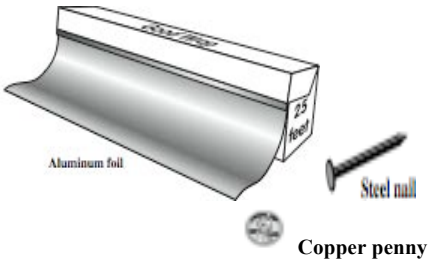
> When an electric fan is running, **most** of the incoming electrical energy changes into which kind of energy?
>
> A. heat energy
> B. light energy
> C. mechanical energy
> D. sound energy

We believe that ELLs do better on items with these words because technical words are generally taught in school science and because they help students understand the correct context for answering the item.

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 (as shown in Example B) or in the answer choices (as shown in Example C).

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

Visuals can be an important source of non-linguistic information for test-takers, particularly ELLs.

Four 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 information about the criteria the student should use is not always provided by the item.
- These items often use a verb, noun, or adjective in conjunction with the extreme value word that can have multiple meanings depending upon the context of use, such as *respond, help, explain, important, and effort* (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 students 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 the assumption of shared background knowledge that ELLs may not always share, 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. For example, the verb "respond" in the earthworm item 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 intended by the writer of Example D.

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.

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).

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.

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.

3. Context-specific Words can be difficult for ELLs because the exact meanings of these words depend on the context in which they appear. These words include:

- high frequency words with multiple meanings (e.g., type)
- words that have relative meanings (e.g., high, low)
- words that can appear as nouns or verbs (e.g., process, form, group) as in: Where did this rock most likely form? vs. Which form of energy toasted the bread?
- non-specialized academic words (e.g., describe, respond, occur).

The non-specialized academic word *respond* posed particular difficulty for the 5th grade ELLs we interviewed, because they invoke one family of meanings in a testing context (e.g., open-response questions) and another family of meanings in a science context (e.g., see the question in Example D, How did the earthworm most likely respond to these conditions?). ELLs often used the testing context to interpret the word *respond*, reading it as referring to their own activity of answering the test question, rather than the behavior of an organism such as an earthworm in response to an environmental stimulus.

Other context-specific words that appeared frequently on the Grade 5 STE MCAS include: *cause, change, characteristic, cycle, damage, explain, form, group, high (higher, highest), in order to, low (lower, lowest), occur, process, produce, represent, result, statement,* and *type*.

4. Low Frequency Non-Technical (LFNT) Words are words that do not appear routinely in 5th grade texts, are not thought of as "scientific", and therefore may not be taught in science class. As a result, ELLs may not know their meanings or how to interpret their use in science test items. In Example G below, LFNT words are highlighted in red.

Examples of LFNT words that occurred most frequently in multiple-choice items on the Grade 5 STE MCAS from 2004-2010 are: *adapted, alligator, approximate, battery, burrowing, classified, conduct, consumers, crystal, drought, erupts, formation, hail, humidity, incomplete, landslide, migration, predators, producer, rotation, slow-moving, sunflower, toaster, topaz,* and *tuna.*

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

Tips For Preparing ELLs for the STE MCAS Test

When reviewing an MCAS item, it is important to find out how students are interpreting the language of the item as it relates to the item's science content.

Students can learn a lot from each other, so provide opportunities for students to think–pair–share and talk with each other about how they made sense of an item and why they selected their answer choice. If you listen to what students are saying, these conversations can also give you valuable insight into students' thinking and highlight those areas that may pose problems for them within the testing situation.

Based on our other work (Rosebery & Warren, 2008[3]) we recommend the following best practices for teaching ELLs. These strategies are good for any age/grade level and subject matter (not just test prep):

- Organize your lessons so students can work in groups
- Allow students frequent opportunities to talk to one another about their understandings of items, texts, etc. so they can learn from one another
- Use strategies like think, pair, share to give students time and an audience to formulate their thoughts and plan out what they are going to say before calling on them in class.
- Hold class discussions so that students can make their thinking visible to one another and to you.

The list below offers some specific suggestions about better preparing students to respond to each of the linguistic feature of science test items.

Forced Comparison Feature

- Examine 3-4 examples of this type of item with students (Additional items with the Forced Comparison feature appear on pages 13-14.). Ask the students what they think the question is asking. Referring to their comments, talk with them about what the meanings of *most likely*, *best*, *greatest, etc.* are in these items. Discuss how they knew what these words meant in the contexts of the items or why they found them confusing. As a group, collectively discuss how they judge what qualifies as the *most likely* or *best* thing in the contexts of the items. Offer alternative words in English or their first languages that mean the same thing (e.g. "most likely" means probably). Suggest and model strategies of trying to think like a scientist when making these judgments. Model asking helpful questions like the following when answering this type of item: 1) What kind of science is the item asking about? 2) What did you learn in science class about this topic?

[3] Rosebery, A. & Warren, B. (2008). Teaching science to English language learners: Building on students' strengths. Arlington, VA: NSTA Press.

Reference Back Feature

- Using an approach similar to the one outlined for Forced Comparison, discuss items that contain the Reference Back feature (Additional items with the Reference Back feature appear on pages 13-14.). Ask students to figure out what information is being referred back to and how to use this information to answer the question.

Technical Language

- Continue to teach ELLs the technical language of science within a unit of study. This helps students to comprehend and correctly answer test items that contain technical words.

Low Frequency Non-Technical (LFNT) and Context-Specific Words

- Vocabulary is best taught within context. Words have different meanings in different contexts and memorizing terms isolated from context is not helpful. Instead, find opportunities within your ongoing science teaching to discuss with ELL students the varied meanings of difficult vocabulary words that arise during lessons, including LFNT words and Context-Specific words.

Visuals

- The use of visuals is an important resource for ELLs. Students need a strong understanding of the different ways visuals are used to convey information. Review and discuss the information found in pictures, tables, charts, and diagrams with students. Provide opportunities for students to make their thinking visible to one other and to you.

- Have conversations with your students about their interpretations of the visual images included in test items, including visual images with arrows of different kinds, pictorial images of objects and processes, and tables. Discuss what they think each visual image is meant to represent in the context of the item, and why they think this.

A Note About Using These Tips:

- Our research shows that when these features appear in STE MCAS items, they can make it harder for ELLs to understand the items. As a result, students are more likely to choose the wrong answer, not because they don't know the science but because they don't understand the English being used. One possible response to these findings is that some teachers will decide to teach students about these features. Alternatively, other teachers may feel that considerable instructional time is already spent on test preparation and may hesitate to expand its scope. We suggest that you weigh these tensions in light of the particular students you are teaching in a given year to make an informed decision.

Tips For Best Practices in Creating Classroom Assessments for ELLs

It can be challenging to create classroom assessments for ELLs that result in accurate reflections of content knowledge. The following tips are intended to help you to modify existing assessments or to create new classroom assessments that will better measure what your ELL students know about science.

Note: If you decide to modify existing test items, review them afterward to make sure you haven't inadvertently deleted key science content words, which might affect the content knowledge being tested.

Technical Language

- Continue using technical language in both the classroom lessons and assessment items you create. During instruction, specifically focus students' attention on the contextual nature of the meanings of technical words within the topic of study.

Visuals

- Continue using items that contain visuals in either the text of the item or the answer choices. During instruction, specifically focus on the various ways visuals are used to support and convey the information being covered.

- When the meaning of a word can be clarified through the use of a visual, include clear, widely recognizable visuals during instruction and in assessment items. Visuals are likely to be especially important for clarifying the meanings of LFNT words that might be unfamiliar to ELLs (e.g. sundial, tulip) and Context-Specific words that might have more than one definition (e.g. outlet, hive).

- Visuals are especially helpful when they illustrate the answer choices. (See Example C. *Flexible Object*, page 5)

- Visuals that represent static objects or organisms appear to be more helpful than visuals that represent processes or procedures that take place over time (i.e. the water cycle, metamorphoses/life cycles, science experiments), which can be open to more than one interpretation.

Reference Back feature

- When creating assessment items and instructional documents, keep the language simple and direct. Avoid using the reference back feature and make the question as explicit as possible.

Forced Comparison Feature

- Try to avoid using the Forced Comparison feature in assessment items. However, students are likely to meet this feature in future standardized testing situations, so provide opportunities for the class to discuss and interpret these types of items together. Model strategies for interpreting "extreme value" terms such as *most likely*, *best*, and *greatest*.

Low Frequency Non-Technical (LFNT) and Context-Specific Words

- Try to avoid the use of LFNT and Context-Specific words, unless accompanied by a visual or other explanation.

Summary

Include	Avoid
Technical Words	Forced Comparison Feature
Visuals	Reference Back Feature
	Context-Specific Words
	Low Frequency Non-Technical Words

Additional MCAS Science Test Items
Containing the Forced Comparison and Reference Back Features

Released items retrieved from Massachusetts Department of Elementary and Secondary Education web site (http://www.doe.mass.edu/mcas/testitems.html).

2010

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

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

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

2009

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

Summary of Results for Three Studies

Study 1, Correlation Study: In Study 1, we investigated the relationship between the presence of the six 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. 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.

Table 1. Effects of Linguistic Features on the Performance of English Language Learners on 5th Grade STE 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 Limited English Proficient (LEP) by their schools, and ten[4] of these students were classified as Formerly LEP 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 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

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

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 5th Grade 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 (ELLs and non-ELLs) 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 ELLs' 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-ELLs who scored below proficient on the MCAS English Language Arts (ELA) test, 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 ELLs, 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 5th Grade 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

For Those Who Would Like to Learn More

Please contact Tracy Noble at Tracy_Noble@terc.edu for further information about any of these studies or for information about how to get copies of any of the following manuscripts:

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., 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., 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