Smarter Balanced Embedded Field Test

For Spring 2015 – updated 7 January 2015

Introduction

Smarter Balanced will embed approximately 6,000 field-test items in its spring 2015 operational English language arts (ELA) assessment and approximately 5,000 field-test items in its mathematics assessment. This document describes the planned implementation of the embedded field test in each subject.

Items will be administered with the following rules:

- On both assessments, embedded field-test (EFT) items may appear at any position between at or after the fifth item on the test and at or before the fifth-from-last item on the test.
- Within the allowable field-test positions, each item or group will be administered in randomly selected positions.
- Item groups (such as items following a passage) will be administered intact.
- The number of field-test items administered to individual students will never exceed the intended maximum nor fall short of the intended minimum.

A subset of the field-test items have been annotated or augmented to support certain accommodations—including, for example, braille, second-language glossaries, and American Sign Language. The customized item-pool filters described in Cohen and Albright (2014) will ensure that students receiving such accommodations will also receive EFT items supporting the same accommodations.

Mathematics Embedded Field Test

The mathematics items to be field tested are all independent, stand-alone items. Each student will be administered exactly two field-test items, embedded in the allowable field-test positions. Each item, therefore, will have approximately a 0.04% exposure rate, yielding an expected 2,400 responses if 6,000,000 students complete the operational mathematics assessment.

ELA Embedded Field Test

While the design for the mathematics assessment is straightforward, the ELA assessment poses more challenges, including the following:

- Most items are embedded in groups, and those groups vary in size.
- Each stimulus will appear with multiple blocks of items.
- The time it takes to answer an item group is not proportional to the number of items but rather depends more heavily on the type of stimulus.

The final bullet is particularly important for writing tasks. These single-item groups include a stimulus and brief writing task that are likely to be more burdensome than a typical stand-alone item.

As described below, each student will see a **minimum of three and a maximum of six** EFT items. Reading sets of items will be constructed with a minimum of three associated items. With this construction, any reading passage will satisfy the minimum requirement and prevent further selections, thereby ensuring that no student receives more than one field-test reading passage.

This section describes

- the plan for constructing EFT blocks;
- the number of EFT items to be administered to each student;
- the implications of this design; and
- next steps that must be accomplished to implement the design.

Plan for Field-Test Blocks

The number of items operationally administered with each stimulus type is constant across grades; the number of stimuli and items developed differs across grades. Our plans for assembling the items into blocks for field testing are summarized in Table 1. The field-test algorithm will ensure that students who receive one block associated with a stimulus do not receive the other block. The system, however, is limited to two blocks per stimulus.

Table 1: Field-Test Blocks to Be Assembled

Item Group	Blocks per Stimulus	Minimum Block Size	Maximum Block Size
Claim 2, Stand-Alone	1	1	1
Claim 3	2	2	3
Claim 1, Lit Short	2	4	4
Claim 1, Info Short	2	4	4
Claim 1, Lit Long	2	5	6
Claim 1, Info Long	2	5	6

Recommended Administration Design and Implications

The field-test algorithm accepts a minimum and maximum number of items to be administered to each student. When the time comes to administer a field-test item, an item group is randomly chosen. Subsequent selections of a field-test items are also random, but if the number of items in the group exceeds the number of remaining "slots," the group is skipped over and another selection is made. If the algorithm cycles through the entire pool without an acceptable selection, no further field-test selections are made.

The algorithm also stops making field-test selections if the minimum number of field-test items has been administered.

AIR recommends setting the minimum number of items to be administered to three and the maximum to six. The slightly lower minimum substantially reduces the number of students who will be asked to respond to multiple writing (Claim 2) prompts.

Table 2 summarizes (1) the expected percentage of students who will see more than one writing task and (2) the percentage of students who will see one, two, three, and four item groups. Table 2 indicates that very few students will see multiple writing tasks, and fewer than one in 5,000 will see more than two. Most students will respond to items in two groups, with fewer than 10%responding to more than two groups of items.

Table 2: Expected Percentage of Students Seeing Multiple Writing Tasks and the Proportion of Students Seeing Each Number of Item Groups, by Grade

| | | Grade | | | | | | |
	Number	3	4	5	6	7	8	HS
Percentage of Examinees Seeing This Many Writing Tasks	**0 or 1**	99.57%	99.27%	99.35%	99.34%	99.34%	99.33%	99.97%
	2	0.43%	0.72%	0.65%	0.66%	0.66%	0.66%	0.03%
	3	0.00%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%
	4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Percentage of Examinees Seeing Each Number of Item Groups	**1**	80.98%	74.60%	75.32%	75.84%	75.84%	75.58%	90.68%
	2	15.52%	20.51%	19.83%	19.49%	19.49%	19.83%	8.06%
	3	3.51%	4.88%	4.85%	4.67%	4.67%	4.59%	1.26%
	4	0.26%	0.26%	0.26%	0.26%	0.42%	0.37%	0.04%

Item exposure rates vary across grades because of different numbers of item groups in the pools. Expected item exposure rates vary within grade as well. This variation happens because of the differences in the number of items in the group. The algorithm will not administer an item group if it will cause the specified maximum number of field-test items to be exceeded. Hence, for example, passages associated with six items are only administered when they are selected as the first item group, whereas stand-alone items may be selected as long as there are any allowable slots left.

Table 3: Item Exposure Rates

	Item Group Type (Size)	Grade						
		3	4	5	6	7	8	HS
Exposure Rates	Claim 2, Stand-Alone	1.10%	1.32%	1.25%	1.25%	1.25%	1.26%	0.34%
	Claim 3 (2 Items)	1.10%	1.16%	1.17%	1.16%	1.16%	1.17%	0.32%
	Claim 3 (3 Items)	1.10%	1.16%	1.17%	1.16%	1.16%	1.17%	0.32%
	Claim 1, Lit Short (4 Items)	1.04%	1.08%	1.09%	1.09%	1.09%	1.10%	0.31%
	Claim 1, Info Short (4 Items)	1.04%	1.08%	1.09%	1.09%	1.09%	1.10%	0.31%
	Claim 1, Lit Long (5 Items)	0.98%	0.99%	0.99%	0.97%	0.97%	0.99%	0.29%
	Claim 1, Lit Long (6 Items)	0.86%	0.85%	0.86%	0.85%	0.85%	0.86%	0.28%
	Claim 1, Info Long (5 Items)	0.98%	0.99%	0.99%	0.97%	0.97%	0.99%	0.29%
	Claim 1, Info Long (6 Items)	0.86%	0.85%	0.86%	0.85%	0.85%	0.86%	0.28%

For the most part, the item exposure rates are approximately 1–2%. The worst case is just under 0.28% in the high school assessment, where a larger pool is being field tested. Assuming a target of approximately 1,500 responses per item, this percentage implies an operational administration to at least 536,000 students in high school. This plan leaves little risk that we will not receive a sufficient number of responses on each item.