Smarter Balanced Scoring Specification

2014–2015 Administration

Summative and Interim Assessments:
ELA Grades 3–8, 11
Mathematics Grades 3–8, 11

Version 2

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

TABLE OF CONTENTS

LIST OF TABLES

1. INTRODUCTION

This document mainly describes the scoring methods of the Smarter Balanced summative assessments designed for accountability purposes during the 2014–2015 test administration. Table 1 lists all summative assessments administered in 2014–2015. Scoring rules for all interim tests are provided at the end of the document.

Table 1. Assessments Administered in 2014–2015

Subject and Grade	Online Administration		Paper Administration	
	Equating Mode	Overall Scoring	Equating Mode	Overall Scoring
ELA 3–8, 11	Pre	MLE	Post	MLE
Math 3–8, 11	Pre	MLE	Post	MLE

Note: MLE = maximum likelihood estimation

2. RULES FOR ESTIMATING STUDENT ABILITY

2.1 Maximum Likelihood Estimation Theta Score

The maximum likelihood estimation (MLE) is used to construct the theta score. Indexing items by *i*, the likelihood function based on the *j*th person's score pattern for k_i items is

$$L_j(\theta \mid \mathrm{z_j}, \mathbf{a}_j, \mathbf{b'}_{1,j}, \ldots \mathbf{b'}_{k_j,j}) = \prod_{i=1}^{k_i} p_i(z_{ji} \mid \theta, a_{i,j}, b_{i,1}^j, \ldots, b_{i,m_i^j}^j),$$

where $\mathbf{b'}_i^j = (b_{i,1}^j, \ldots, b_{i,m_i^j}^j)$ are the *i*th item's step parameters and m_i^j is the possible score of this item, $a_{i,j}$ is the discrimination parameter. Depending on the item type, the probability $p_i(z_{ji} \mid \theta, a_{i,j}, b_{i,1}^j, \ldots, b_{i,m_i^j}^j)$ takes either the form of a two-parameter logistic (2PL) model for multiple-choice (MC) items or the form based on the generalized partial credit model for polytomous items.

In the case of MC items, we have

$$p_i(z_{ji} \mid \theta, a_{i,j}, b_{i,1}^j, \ldots, b_{i,m_i^j}^j) = \begin{cases} \dfrac{\exp Da_{i,j}(\theta - b_i^j)}{1 + \exp Da_{i,j}(\theta - b_i^j)} = p_i & \text{if } z_{ji} = 1 \\ \dfrac{1}{1 + \exp Da_{i,j}(\theta - b_i^j)} = 1 - p_i & \text{if } z_{ji} = 0 \end{cases};$$

in the case of constructed-response (CR) items,

$$p_i(z_{ji} \mid \theta, a_{i,j}, b_{i,1}^j, \ldots, b_{i,m_i^j}^j) = \begin{cases} \dfrac{\exp Da_{i,j}(\sum_{r=1}^{z_{ji}}(\theta - b_{i,r}^j))}{s_i(\theta, b_{i,1}^j, \ldots, b_{i,m_i^j}^j)} & if\ z_{ji} > 0 \\ \dfrac{1}{s_i(\theta, b_{i,1}^j, \ldots, b_{i,m_i^j}^j)} & if\ z_{ji} = 0 \end{cases},$$

where $s_i(\theta, a_{i,j}, b_{i,1}^j, \ldots, b_{i,m_i^j}^j) = 1 + \sum_{l=1}^{m_i^j} \exp(\sum_{r=1}^{l} Da_{i,j}(\theta - b_{i,r}^j))$, $D = 1.7$.

Thus, we have $SE\left(\hat{\theta}\right) = \sqrt{Var\left(\hat{\theta}\right)}$.

2.2 Scoring All Correct and All Incorrect Cases

In item response theory (IRT) maximum likelihood (ML) ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. To handle such cases, AIR proposed several options. The method below has been agreed on by both Smarter Balanced and AIR.

For all correct and all incorrect cases, assign the highest obtainable scores (HOT and HOSS) or the lowest obtainable scores (LOT and LOSS) presented in Table 3.

3. SCORING INCOMPLETE TESTS

3.1 Overview

Sometimes students fail to complete their tests. This section covers three specifications:

- When a test is considered attempted
- When a test is scored
- How incomplete tests are scored when they are scored

3.1.1 Attemptedness/Participation

All items are included in the evaluation of test records for attemptedness, including the following:

- Operational and field test items
- All item types

All tests with at least one item answered are considered attempted, and will receive a score.

Note that different states may implement different rules around this, and those rules will influence what gets reported.

3.1.2 When to Score an Incomplete Test

All attempted tests get scored if the tests meet the rules of attemptedness defined in Section 3.1.1.

3.1.3 Assigning Scores to Incomplete Tests

Tests are considered "complete" if students respond to the minimum number of operational items specified in the blueprint for both the CAT and the performance task. Otherwise, the tests are "incomplete." MLE is used to score the incomplete tests counting unanswered items as incorrect. If a student completes a test, but did not submit the test, TDS marks the test as completed.

3.1.3.1 Adaptive Tests (CAT segment)

When tests are adaptive, the specific unanswered items are unknown; thus, simulated items are used in place of administered items. Simulated items are generated with the following rules:

- Minimum of operational test length is used to determine the test length of the incomplete tests.
- It is assumed that all unanswered operational items are MC items. The item parameters of all unanswered operational items are equal to the average values of the answered operational items for discrimination and difficulty parameters, respectively.
- All unanswered operational items are scored as "incorrect."
- Because the content standards for the unanswered operational items are unknown, only an overall performance (total score) will be computed, but not subscores (claim scores).
- If the responded items are all incorrect, assign the lowest obtainable scores (LOT and LOSS) in Table 3.

3.1.3.2 Fixed Form Tests

For fixed form tests, including the paper summative tests and performance tasks, unanswered items will be treated as incorrect. For summative fixed form tests, both total and subscores will be computed.

3.1.3.3 Merging Online and Paper Tests

This testing program provides both online and paper tests. Therefore, there will be cases where a student takes part of the test online and part of the test on paper. For these tests, the items administered online and paper will be combined before generating total and subscores. In some cases, a student will take the same part (performance task or CAT) online and on paper. If one version is complete and the other is not complete, the complete version will be chosen. If multiple incomplete tests exist, the most complete test will be used. Otherwise, the online version will always be chosen for scoring purposes. No attempt will be made to merge multiple

incomplete attempts into a single test event. If multiple complete tests with the same administration mode exist for a student, the most recent version will be used.

3.2 Attemptedness for Reporting

All tests with at least one operational item answered in the CAT segment and the performance task segment, respective, are considered attempted and will receive a score.

Attemptedness rules for CAT items:

- N (not attempted) = responded to zero item
- Y (attempted) = responded to one item or more

Attemptedness rules for performance task items:

- N (not attempted) = responded to zero item
- Y (attempted) = responded to one item or more

Report an overall score the following occurs:

- CAT attemptedness = Y; and
- Performance task attemptedness = Y

Note that different states may implement different rules around this, and those rules will influence what gets reported.

3.3 Scoring Rules

3.3.1 Performance Task Scoring Rules

Scoring rules for performance task items:

- Any condition code will be recoded to zero.
- Evidence, purpose, and conventions are the scoring dimensions for the writing essays. Scores for evidence and purpose dimensions will be averaged, and the average will be rounded up.

3.3.2 Hand Scoring Rules

3.4 Reporting Rules

A total and subscores will be reported if both CAT and the performance task portion of the test are attempted.

4. RULES FOR TRANSFORMING THETA TO VERTICAL SCALE SCORES

The IRT vertical scale is formed by linking across grades using common items in adjacent grades. The vertical scale score is the linear transformation of the post-vertically scaled IRT ability estimate.

$$SS = a * \theta + b$$

The scaling constants *a* and *b* are provided by Smarter Balanced. Table 2 lists the scaling constants for each subject for the theta-to-scaled score linear transformation. Scale scores will be rounded to an integer.

Table 2. Vertical Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA	3-8, HS	85.8	2508.2
Math	3-8, HS	79.3	2514.9

4.1 Lowest/Highest Obtainable Scores

Extreme unreliable student ability estimates will be truncated. Table 3 presents the lowest and the highest obtainable scores in both theta and scale score metrics. Estimated theta's lower than LOT or higher than HOT will be truncated to the LOT and HOT values, and assign LOSS and HOSS associated with the LOT and HOT.

The standard error for LOT and HOT will be computed using the LOT and HOT ability estimates given the administered items. For example, in the formula in Section 5.1, $\hat{\theta}$=LOT or HOT, a and b are for the administered items, and $SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}$. SE(SS)= $SE(\hat{\theta}) * slope$.

Table 3. Lowest and Highest Obtainable Scores

Subject	Grade	Theta Metric		Scale Score Metric	
		LOT	HOT	LOSS	HOSS
ELA	3	-4.5941	1.3374	2114	2623
ELA	4	-4.3962	1.8014	2131	2663
ELA	5	-3.5763	2.2498	2201	2701
ELA	6	-3.4785	2.5140	2210	2724
ELA	7	-2.9114	2.7547	2258	2745
ELA	8	-2.5677	3.0430	2288	2769
ELA	HS	-2.4375	3.3392	2299	2795
Math	3	-4.1132	1.3335	2189	2621
Math	4	-3.9204	1.8191	2204	2659

Math	5	-3.7276	2.3290	2219	2700
Math	6	-3.5348	2.9455	2235	2748
Math	7	-3.3420	3.3238	2250	2778
Math	8	-3.1492	3.6254	2265	2802
Math	HS	-2.9564	4.3804	2280	2862

5. CALCULATING MEASUREMENT ERROR

5.1 Standard Error of Measurement

With MLE estimation, the standard error (SE) for student *i* is:

$$SE(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}}$$

where $I(\theta_i)$ is the test information for student *i*, calculated as:

$$I(\theta_i) = \sum_{j=1}^{I} D^2 a_j^2 \left(\frac{\sum_{l=1}^{m_j} l^2 Exp\left(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk})\right)}{1 + \sum_{l=1}^{m_j} Exp\left(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk})\right)} - \left(\frac{\sum_{l=1}^{m_j} lExp\left(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk})\right)}{1 + \sum_{l=1}^{m_j} Exp\left(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk})\right)} \right)^2 \right)$$

where $m_j$is the maximum possible score point (starting from 0) for the *j*th item, *D* is the scale factor, 1.7.

SE is calculated based only on the answered item(s) for both complete and incomplete tests. The upper bound of SE is set to 2.5 on theta metric. Any value larger than 2.5 is truncated at 2.5 on theta metric.

5.2 Standard Error Transformation

Standard errors of the MLEs are transformed to be placed onto the reporting scale. This transformation is:

$$SE_{vs} = a * SE_{\theta_i}$$

where SE_θ is the standard error of the ability estimate on the θ scale and a is the slope of the scaling constants that take θ to the reporting scale.

6. RULES FOR CALCULATING CLAIM SCORES (SUBSCORES)

6.1 MLE Scoring for Claim Scores

Claim scores will be calculated using MLE, as described in Section 2.1; however, the scores are based on the items contained in a particular claim.

6.2 Scoring All Correct and All Incorrect Cases

Apply the rule in Section 2.2 to each Claim.

6.3 Reporting Strengths and Weaknesses at Claim Level for Each Student

AIR will report relative strengths and weaknesses for each student at the reporting category (claim) level in addition to scaled scores. If the difference between the proficiency cut score and the claim score is greater (or less) than 1.5 standard error of the claim, a plus or minus indicator will appear on the student's score report. The specific rules are as follows:

- $\theta_{rc} - \theta_p \leq -1.5SE_{rc}$ indicates performance below proficient
- $\theta_{rc} - \theta_p \geq 1.5SE_{rc}$ indicates performance above proficient
- Otherwise = 0, a strength or weakness is indeterminable

where θ_{rc} is the student's score on a reporting category, and θ_p is the proficiency theta cut, and se_{rc} is the standard error for the student's score on the reporting category.

7. RULES FOR CALCULATING PERFORMANCE LEVELS

Overall scale scores for Smarter Balanced are mapped into four performance levels per grade/course. The performance level designations are level 1, level 2, level 3, and level 4. The definition of these levels is defined after standard setting.

7.1 Threshold Scale Scores for Four Achievement Levels

Tables 4 and 5 show the theta cut scores and reported scaled scores (SS) for the ELA assessments and the math assessments, respectively.

Table 4. ELA Theta Cut Scores and Reported Scaled Scores

Grade	Theta Cut between Levels 1 and 2	SS Cut between Levels 1 and 2	Theta Cut between Levels 2 and 3	SS Cut between Levels 2 and 3	Theta Cut between Levels 3 and 4	SS Cut between Levels 3 and 4
3	-1.646	2367	-0.888	2432	-0.212	2490
4	-1.075	2416	-0.410	2473	0.289	2533
5	-0.772	2442	-0.072	2502	0.860	2582
6	-0.597	2457	0.266	2531	1.280	2618
7	-0.340	2479	0.510	2552	1.641	2649

8	-0.247	2487	0.685	2567	1.862	2668
HS	-0.177	2493	0.872	2583	2.026	2682

Table 5. Math Theta Cut Scores and Reported Scaled Scores

Grade	Theta Cut between Levels 1 and 2	SS Cut between Levels 1 and 2	Theta Cut between Levels 2 and 3	SS Cut between Levels 2 and 3	Theta Cut between Levels 3 and 4	SS Cut between Levels 3 and 4
3	-1.689	2381	-0.995	2436	-0.175	2501
4	-1.310	2411	-0.377	2485	0.430	2549
5	-0.755	2455	0.165	2528	0.808	2579
6	-0.528	2473	0.468	2552	1.199	2610
7	-0.390	2484	0.657	2567	1.515	2635
8	-0.137	2504	0.897	2586	1.741	2653
HS	0.354	2543	1.426	2628	2.561	2718

8. RULES FOR INTERIM TESTS

This year all interim tests are fixed-form tests. Interim ICAs will be scored in the same way as the summative tests. Interim Blocks will each receive an overall score and a proficiency classification (the same calculation as claim scores).