Smarter Balanced Scoring Specifications

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

Updated July 28, 2016

TABLE OF CONTENTS

LIST OF TABLES

1. INTRODUCTION

This document describes the scoring methods of the Smarter Balanced summative assessments designed for accountability purposes during the 2015–16 test administration. Table 1 lists all summative assessments administered in 2015–16. In some instances, the document specifies options available to vendors that may differ from the approach used in the open source test scoring system.

Table 1. Assessments Administered in 2015–16

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

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

where $\mathbf{b}'^j_i = (b_{i,1}^j, \ldots, b_{i,m_i^j}^j)$ 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,

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

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

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 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 4.

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

If a student logged onto both the CAT and the Performance Task parts of the test, the student is considered as participated, even if no items are answered. These tests will be included in the data file, but no scores will be computed.

3.1.2 When to Score an Incomplete Test

All attempted tests get scored if the tests meet the rules of attemptedness.
All tests with at least one CAT item and one performance task item answered are considered attempted. For the interim assessment blocks (IABs), a block with at least one item answered is considered attempted.

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

Attemptedness rules for Block items (IAB):

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

For Summative and ICA, report scores the following occurs:

- CAT (non-performance task part) attemptedness = Y; AND
- Performance task attemptedness = Y

For Interim Assessment Blocks (IABs), report scores the following occurs:

- Block attemptedness = Y

Attemptedness Flag in the data file

The attemptedness flag will include three values for Summative and ICA (N, P, and Y) and two values (P and Y) for IAB.

N = non-participant (a student who only had activity on a single part of the test – CAT or PT, but not both)

P = participant (a student who logged into both parts of the test but didn't respond to anything on at least one part of the test)

Y = attempted (a student who logged into both parts of the test and responded to at least one item on both)

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 the CAT and all items in the performance task form. 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. If TDS allowed the student to submit his/her test it will be considered "complete".

3.1.3.1 Online Summative Tests

Online Summative Tests include both the CAT and the performance task parts. The performance task part includes a fixed form test. For the performance task items, unanswered items will be treated as incorrect.

For the CAT items, the identity of most of the specific unanswered items are unknown; If items have been lined up for administration (through the pre-fetch process), parameters are known and the items are scored as incorrect. That is, they are treated in the same manner as known items in interim tests, paper/pencil tests and performance tasks. For the remainder of items, simulated parameters are used in place of administered items. In the open source scoring system, simulated item parameters are generated with the following rules:

- Minimum of the CAT 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 on-grade items for discrimination and difficulty parameters in the summative item pool, respectively. See Table 2 for the average discrimination and difficulty parameters.
- All unanswered operational items are scored as "incorrect."

Table 2. Average Discrimination (a) and Difficulty (b) Parameters

Grade	ELA		Math	
	a	b	a	b
3	0.67	-0.42	0.85	-0.81
4	0.59	0.13	0.81	-0.06
5	0.61	0.51	0.77	0.68
6	0.54	1.01	0.70	1.06
7	0.54	1.11	0.71	1.79
8	0.53	1.30	0.61	2.29
HS	0.50	1.69	0.53	2.71

Vendors may use other equivalent methods of generating item parameters (e.g., inverse TCC [Stocking, 1996]). For the summative online test, if the CAT part is incomplete, only a total score will be reported, but not subscores because the claim information for the unanswered CAT items is unknown.

3.1.3.2 Fixed Form Tests

For fixed form tests, including the paper summative tests, ICAs and IABs, unanswered items will be treated as incorrect. For summative fixed form tests and ICAs, 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 Hand Scoring Rules

Scoring rules for hand scoring 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 Reporting Rules

Scores will be reported for all tests that meet the attemptedness rule in Section 3.1.2.

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 3 lists the scaling constants for each subject for the theta-to-scaled score linear transformation. Scale scores will be rounded to an integer.

Table 3. 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 Scale Scores (HOSS/LOSS)

HOSS/LOSS Options

Options for HOSS/LOSS values have been set in policy. Implementation of the option desired by each member needs to be negotiated with the test scoring contractor. In 2015-16 Smarter Balanced members have the following options:

Option 1: Members may choose to retain the 2014-15 LOSS/HOSS values which are shown in Table 4.

NOTE: For 2015-16, the Smarter Balanced open source test delivery system retained the 2014-15 LOSS/HOSS values and used the scoring rules described in the 2014-15 Scoring Specifications.

Table 4. 2014 – 2015 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

Option 2: Members may choose to use other LOSS/HOSS values beginning in SY 15/16 as long as the revised LOSS values do not result in more than 2% of students falling below the LOSS level and the revised HOSS values do not result in more than 2% of students falling above the HOSS level.

Option 3: Members may choose to eliminate LOSS/HOSS altogether.

Additional Considerations

- All-wrong/All-right tests:
    - For all incorrect tests, score by adding 0.5 to an item score with smallest a-parameter among the administered operational items (CAT and PT) for a test.
    - For all correct cases, score by subtracting 0.5 from an item score with smallest a-parameter among the administered operational items (CAT+PT) for a student.
- Smarter Balanced will need to retain both the calculated theta score and the reported scale score for students whose scores fall into HOSS/LOSS ranges.
- If using Option #1 or #2 above:

- When the scale score corresponding to the estimated theta is lower than LOSS or higher than HOSS, the scale score will be assigned associated LOSS and HOSS values. The theta score will be retained as originally computed.
- LOSS and HOSS scale score rules will be applied to all tests (Summative, ICA, and IAB) and all scores (total and subscores).

- The standard error for LOSS and HOSS will be computed using theta ability estimates given the administered items. For example, in the formula in Section 5.1, $\hat{\theta}$=theta for the LOSS or HOSS, a and b are for the administered items.
- If using Option #3, the scale score is calculated directly from estimated theta.

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 transform 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.

In ELA, claim scores will be computed for each claim. In math, claim scores will be computed for Claim 1, Claim 2 and 4 combined, and Claim 3.

6.2 Scoring All Correct and All Incorrect Cases

Apply the rule in Section 2.2 to each Claim.

6.3 Rules for Calculating Strengths and Weaknesses for Claims (Reporting Categories)

Relative strengths and weaknesses for each student at the reporting category (claim) level are reported 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.

For IAB and Summative, the specific rules are as follows:

- Below Standard (Code=1): if $round(SS_{rc} + 1.5 * SE(SS_{rc}),0) < SS_p$
- At/Near Standard (Code=2) : if $round(SS_{rc} + 1.5 * SE(SS_{rc}),0) \geq SS_p$ and $round(SS_{rc} - 1.5 * SE(SS),0) < SS_p$, a strength or weakness is indeterminable
- Above Standard (Code=3): if $round(SS_{rc} - 1.5 * SE(SS_{rc}),0) \geq SS_p$

where SS_{rc} is the student's scale score on a reporting category; SS_p is the proficiency scale score cut (Level 3 cut); and $SE(SS_{rc})$ is the standard error of the student's scale score on the reporting category. Assign *Above Standard* (code=3) to HOSS and assign *Below Standard* (code=1) to LOSS.

For ICA, the rules for calculating achievement levels are as follows:

- Below Standard (Code=1): if $a * (\theta_{rc} + 1.5 * SE(\theta_{rc})) + b < SS_p$
- At/Near Standard (Code=2) : if $[a * (\theta_{rc} + 1.5 * SE(\theta_{rc})) + b] \geq SS_p$ and $[b * (\theta_{rc} - 1.5 * SE(\theta_{rc})) + a] < SS_p$, a strength or weakness is indeterminable
- Above Standard (Code=3): if $[a * (\theta_{rc} - 1.5 * SE(\theta_{rc}) + b]) \geq SS_p$

where θ_{rc} is the student's theta score on a reporting category. SS_p is the proficiency scale score cut (Level 3 cut). $SE(\theta_{rc})$ is the standard error of the student's score on the reporting category. a and b are the scaling constants.

[Note: The difference in the two rules is in the rounding rule. Because a rounding rule was updated after ICA was deployed, ICA has a different rule.]

7. RULES FOR CALCULATING ACHIEVEMENT LEVELS

Overall scale scores for Smarter Balanced are mapped into four achievement levels per grade/content area. The achievement level designations are Level 1, Level 2, Level 3, and Level 4. The definition of these levels was defined after achievement level setting.

7.1 Threshold Scale Scores for Four Achievement Levels

Tables 5 and 6 show the theta cut scores and reported scaled scores (SS) for the ELA/literacy assessments and the mathematics assessments, respectively.

Table 5. ELA/Literacy 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 6. Mathematics 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.

For the Interim Assessment Blocks (IABs), the test results per grade and content area will be merged into a single result, and the block scores will be calculated as reporting category scores on the combined result. At the overall level for the combined result, number of blocks attempted and number of blocks proficient will be computed for reporting purposes in the Online Reporting System as following:

- Number of blocks attempted: Count the Blocks with Block attemptedness=Y
- Number of block proficient: Count the Blocks with performance “Above Starndard”.

In addition, the IAB test results will also be scored individually (independently from the combined test). There will be overall scores on each IAB test (attemptedness, scale score, and proficiency level) that use the same calculation rules (and are the same) as the reporting category scores for the blocks represented in the combined test. But these will be used to produce open source scoring configuration packages and for delivery of results and scores to other clients and vendors.

REFERENCES

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. Journal of Educational and Behavioral Statistics, 21 365-389.

REVISION LOG

Updates to the 2014-2015 Scoring Specifications are noted below.

Section	Page	Description of Change	Revision Date
1	1	Section 1 provides guidance regarding the availability of options to vendors that may differ from the approach used in the open source system.	7/28/16
3.1.3.1	4	Section 3.1.3.1 revised to include information regarding unanswered items and the use of the pre-fetch process. Clarified the use of simulated item "parameters." Added information about vendor use of other equivalent methods of generating item parameters.	7/28/16
4.1	5	Section 4.1 revised to include member options for the use of HOSS/LOSS	7/28/16