Initial Report on the Calibration of Paper and Pencil Forms

UCLA/CRESST – August 2015

This report describes the procedures used in obtaining parameter estimates for items appearing on the 2014-2015 Smarter Balanced Assessment Consortium (SBAC) summative paper-pencil forms. Among the items appearing on these forms, a subset were identified as "modified"—meaning it was assumed that these items would function differently in this mode of administration and that it would not be reasonable to use the item parameters from the previous calibration (in which items were administered via computer). The remaining items—which we refer to in this report as "unmodified"—were treated as candidate anchor items for the purpose of linking the paper-pencil forms to the scale of the computer-based test administrations.

For each grade and subject, our calibration of items for paper and pencil administration included the following steps:

1. Exclusion of ineligible cases. Data for the calibration of items appearing on the paper and pencil forms were provided by two consortium member states. We excluded from the calibrations any cases with fewer than nine valid item scores on the form. Invalid score codes included B=blank, L=non-scorable language, T=off-topic, M=off-purpose. Items with no score or score code were also treated as invalid for the purpose of determining eligibility. Table 1 below shows the number of eligible cases used in the paper-pencil calibrations, by grade level and subject.

Table 1. Calibration sample size by grade and subject.

Grade	ELA	Math
3	3260	2647
4	1922	1956
5	1986	2190
6	1434	1425
7	1561	1422
8	1534	1000
HS	1000	1000

2. Recoding of invalid or missing item scores. For cases meeting the inclusion criterion in step #1 (i.e., having a minimum of nine valid item scores on the form), any items with an invalid score code (or missing a score or score code entirely) were recoded with a score of zero (i.e., treated as incorrect or receiving the minimum score).

3. Initial calibration. We performed an initial calibration in which (a) the parameters of all "unmodified" (candidate anchor) items were fixed to their prior estimates (the values obtained from the online administration), (b) the parameters of all "modified" items were freely estimated, and (c) the latent variable density was estimated as an empirical histogram (see, e.g., Woods, 2007; Houts & Cai, 2013) with estimated mean and variance. This population distribution was estimated due to our uncertainty about the proficiency of the students used in this calibration and the expectation that these students were unlikely to be representative of all students in the grade level. In order to match the scoring procedures used with the computer-based tests, scores for two items were first recoded (two score levels collapsed into one). Another two items required similar recoding in order to obtain stable item parameter estimates. All such recoding was noted in the scoring parameter files created after the final calibration. From the initial calibration performed in

this step, we obtained item and model goodness-of-fit indices. The calibrations in this step (as well as in later steps #4 and #6) were performed with the flexMIRT item response modeling software (Cai, 2015).

4. Calibrations to estimating of parameters of "unmodified" items. We then performed a series of calibrations identical to step #3 but with the parameters of one "unmodified" item at a time now freely estimated. The parameters of all other "unmodified" items were fixed to their prior estimates (from the prior field test, in which items were administered online). As in step #3, the parameters of all "modified" items were freely estimated, along with the population distribution's mean, variance, and shape. The number of calibrations performed in this step was equal to the number of "unmodified" items on the form (for the given subject and grade level). From these calibrations, we obtained item and model goodness-of-fit indices and item parameter estimates for the "unmodified" items when these parameters were estimated freely, rather than fixed to their prior estimates (as in step #3).

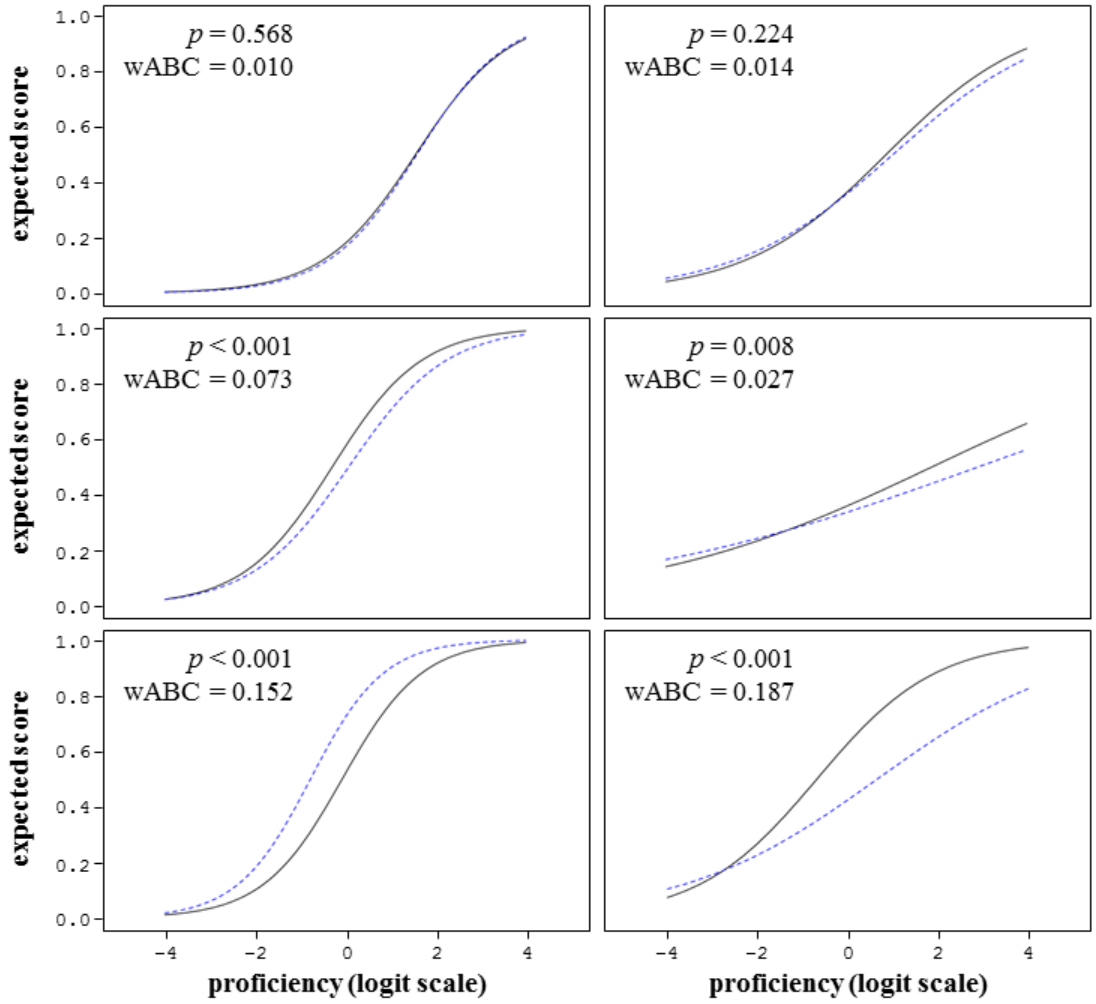
5. Evaluation of candidate anchor items. We then examined the extent to which each "unmodified" item should be retained or rejected as an anchor in the final calibration for the paper and pencil forms (i.e., whether or not it would be reasonable to fix the parameters of these items to their prior estimates). We used the parameter estimates from the prior item calibration (in which the items were administered via computer) and the calibrations from step #4 to compute the expected score functions for the two modes of test administration. For dichotomous items, the expected score function is simply the item traceline. The two expected score functions (for the computer-based and paper-pencil administrations) were plotted, and differences in item functioning across the modes of administration were quantified by computing a weighted Area Between the Curves (wABC; see Hansen, Cai, Stucky, Tucker, Shadel, & Edelen, 2014).

$$\text{wABC}_i = \int_\theta \left| \text{ES}_i^{(\text{C})}(\theta) - \text{ES}_i^{(\text{P})}(\theta) \right| \text{g}(\theta) d\theta.$$

Here, $\text{ES}_i^{(\text{C})}(\theta)$ is the expected score function based on the parameter estimates from the prior calibration, in which items were administered via computer, and $\text{ES}_i^{(\text{P})}(\theta)$ is the expected score function based on the parameters estimated from the paper-pencil calibration (in step #4). The proficiency distribution for the grade and subject is given by $\text{g}(\theta)$. The weighted differences were integrated over the range of proficiency levels (from -4 to +4 on the logit scale, in increments of 0.1).

We rejected as anchors any items with a wABC value greater than 0.150. Note that all items with wABC > 0.150 also showed significant misfit based on the marginal fit index obtained in the initial calibration and significant improvement in overall model fit when the parameters of the item were estimated freely rather than fixed to their prior estimates (evaluated though a likelihood ratio test; note that the model in step #3 is nested within each model estimated in step #4). There were, however, items that displayed statistically significant differences in item functioning (based on the likelihood ratio test) that showed little meaningful difference in the expected score functions (wABC < 0.150); these items were retained as anchors. Examples of the expected score functions and corresponding wABC values are shown in Figure 1, for six Grade 3 ELA items.

Figure 1. Expected score functions for six items, by mode of administration.



Notes: Items shown are from the Grade 3 ELA paper-pencil form. Solid black lines show the expected score functions based on the parameter estimates obtained from the prior calibration (when items were administered via computer). Dotted blue lines show the expected score functions when the item's parameters were estimated, rather than fixed. The *p*-value is for a likelihood ratio test examining whether the overall fit of model in which the item's parameters are fixed to the previous estimates (i.e., the model estimated in step #3) is significantly worse than a model in which the item's parameters are freely estimated (as in step #4). wABC is the weighted difference in expected score functions.

For the two items in the first row, model fit was not significantly worse when the item parameters were fixed (i.e., $p > 0.05$ for the likelihood ratio test comparing the models from steps #3 and #4). For the two items in the second row, model fit was significantly worse when the parameters were fixed, but the wABC value remained below the threshold value of 0.150. For items in the third row, model fit was significantly worse when item parameters were fixed to the previous values, and the wABC values exceeded 0.150. Among these six items, only these last two items were rejected as anchors. Table 2 presents a summary of results from our evaluation of the "unmodified" (candidate anchor) items.

Table 2. Number of items in calibration, by grade and subject.

grade	modified items	unmodified items		total
		anchor	rejected	
		ELA		
3	20	30	0	50
4	19	31	0	50
5	14	36	0	50
6	14	37	0	51
7	19	30	2	51
8	22	28	2	52
HS	12	31	3	46
		Math		
3	23	18	0	41
4	23	16	1	40
5	16	24	1	41
6	24	13	1	38
7	25	16	0	41
8	22	17	0	39
HS	19	18	0	37

Notes: Performance task items for the HS ELA and Math forms were completed by only a small proportion of students the sample used in these analyses. As a result, the HS performance task items were excluded from the calibration. Two items in Grade 6 ELA were also excluded from the calibration because very few examinees answered these items correctly, leading to unstable parameter estimates. The excluded items are not included in the item counts shown in this table but are used in scoring the paper-pencil forms (using the parameters from the prior calibration).

The number of "unmodified" items within the form for a particular grade and subject that displayed wABC values exceeding the threshold value of 0.150 ranged from zero to three. For those subjects and grades in which no items were rejected, all "unmodified" items were used as anchors, the model from step #3 was used as the final calibration model, and the parameter estimates from this model were provided to SBAC for use in scoring the paper-pencil forms. When one or more items were rejected as anchors, we proceeded to step #6.

6. Final item calibration. For grades and subjects in which any "unmodified" item was rejected as an anchor, we estimated a final calibration model that was identical to the model from step #3, except that the parameters of all rejected anchor items were freely estimated, rather than fixed to their prior estimates. Parameters of the "modified" items were also freely estimated. As in steps #3 and #4, the latent variable density was estimated as an empirical histogram. The parameter estimates from this final calibration were provided to SBAC for use in scoring the paper-pencil forms.

References

Cai, L. (2015). *flexMIRT: Flexible multilevel item factor analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. S., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine and Tobacco Research, 16, Supplement 3,* S175-S189.

Houts, C. R., & Cai, L. (2013). *flexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and psychological measurement, 67,* 73-87.