Standard Setting for a Common Core Aligned Assessment

National Council on Measurement in Education (NCME)
Philadelphia, PA

Ye Tong, Pearson
Brian Patterson, Pearson
Peter Swerdzewski, Regents Fund
Candace Shyer, New York State Department of Education

April 2014

Abstract

This paper discusses an implementation of Evidence Based Standard Setting (EBSS for) common core aligned assessments in grades 3-8 English language arts and mathematics using external benchmark data from SAT, PSAT, and NAEP. Policy considerations took into account the expectation that students leave high school ready for college or career and data about current levels of college readiness in the state were shown to panelists in support of this intended policy inference. Panelists were additionally shown impact data associated with the external benchmarks and were asked to provide a range of expected impact data for the proficient cut score. Finally, panelists were given a range of bookmark placements for the proficient cut that would align well with these external benchmarks.

Keywords: common core, standard setting,

Standard Setting for a Common Core Aligned Assessment

Since the introduction of the Common Core State Standards (CCSS) in 2010, forty-five states, the District of Columbia, four territories, and the Department of Defense Education Activity (DoDEA) have adopted the CCSS. Two large-scale assessment consortia, Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness in College and Careers (PARCC), were formed to assess the CCSS in their participating states starting in the 2014-2015 school year. Many states also adopted or aligned their standards to the common core. In this paper, we discuss a standard setting we planned and facilitated for Common Core aligned assessments for English Language Arts and Mathematics in grades 3 through 8.

Whenever new standards are introduced, new assessment should be developed to measure the new standards. A standard setting process typically is conducted to set performance standards for the new assessment. The standard setting activities were conducted for this new assessment system to achieve the following goals:

- Provide vertical articulated performance standards for the assessment in ELA and mathematics and indicate the degree to which students have met the standards for their grade
- Incorporate college readiness into the performance standards for the new assessment
- Establish rigorous and attainable performance standards
- Incorporate policy considerations into the established performance standards
- Provide information to improve instruction

With these goals in mind, the process for setting performance standards needed to be expanded from the previous process used to establish performance standards for the existing assessment. Specific extensions to the standard-setting process included:

1. Taking into account the grade-level standards assessed and indicators of college readiness to the degree practicable.
2. Using external benchmarks (e.g., college performance data, SAT scores, etc.) to inform setting performance standards.
3. Incorporating policy and content considerations into the standard setting process
4. In order to achieve the various goals for the standard setting, the following high-level steps were adopted for the standard setting process
5. Set external benchmarks for the assessment using empirical studies
6. Develop and review both policy and grade-level specific performance level descriptors (PLDs)
7. Convene standard-setting committees having representations from both content and policy experts
8. Conduct vertical articulation
9. Policy makers approve the performance standards recommended from the standard setting process

Evidence-based standard setting process (McClarty et al, 2013) was used to help set performance standards for the Common Core aligned assessments. Evidence-based process combines content, data, educator expertise, and policy expertise with a standard-setting procedure:

- *Content:* performance labels, policy definitions, curriculum, assessment blueprints, and specific PLDs
- *Data:* external benchmarks that help empirically define college readiness using assessment test results
- *Educator expertise:* standard-setting committee reviews of content and data
- *Policy expertise:* standard-setting committee reviews of content and data
- *Standard-setting procedure:* use of a standard-setting methodology, the Bookmark (Mitzel et al., 2001), to recommend performance standards

In this paper, we discuss the various aspects of the standard setting, with the emphasis on how empirical data were incorporated into the standard setting process. No specific data will be shared in this paper due to confidentiality concerns. We will focus on the process.

Set External Benchmarks for Assessment Using Empirical Studies

In order to have empirical evidence linking college readiness information onto the assessment, external benchmark studies were conducted. In the following, we describe studies discussed along with their considerations.

Linking Studies with State Universities' Data

To evaluate students' readiness for college, freshman Grade Point Average (GPA) seems to be a direct indicator. Initial discussions centered on some type of linking studies that could examine the relationship between students' freshman year GPA and their high school assessment test scores. Logistical regression was discussed as one of the possible analysis to establish that link: students have to score at this level for the high school assessment in order to be college ready. Such findings could provide direct information on the relationship between how high school assessment results could predict students' performance in their freshman year in college.

While these linking studies appeared to be good indicators, there were also a number of disadvantages we had to consider.

First, through some initial analysis, it was observed that the correlation coefficient between freshman year GPA and high school assessment scores was very low (in the neighborhood of 0.2 to 0.3), partly due to the limited scale from freshman year GPA. Additionally, there is a 4-year lag in collecting the high school (9^{th}-grade) test scores and students' college freshman year GPA. Therefore, historical assessment for high school will have to be used. In this state, the high school assessments had gone through transformation in 2012. As a result, even old assessments, instead of the most recent version based on the current academic standards, would have to be used in the analysis. Thirdly, such analysis will only include students who attended state universities. Therefore, it would only constitute around 30% of the students in the cohort, and may introduce some bias in the results. Lastly, the standard setting was for the grades 3–8 assessments, whereas this study focuses on the relationship between high school test and college readiness. Additional link between the high school assessments and the grades 3–8 assessments would have to be established.

Due to these limitations, such direct link between high school assessments and freshman GPA was discussed but not included in this standard setting process. The State would consider such analysis when performance standards would be set for the high school exams.

Analysis Using the PSAT and SAT

PSAT and SAT are prevalent in the state. The state sponsors all 10^{th} graders to take the PSAT, and the majority of the students take SAT in their 11^{th} grade. Therefore, we discussed using PSAT and SAT related analysis for the external benchmark studies. The College Board has conducted research on the college readiness indicators on the PSAT and SAT (Wiley et al.,

2011). College readiness was defined as having at least a 65 percent probability of obtaining a B- (or 2.67 first year GPA) or higher for the first year freshman courses. The college readiness definition adopted by the State, however, was different from the College Board definition. Therefore, additional analysis was carried out to identify the PSAT and SAT scores associated with the State's college readiness definition.

In this set of the analysis, the empirical definition of college readiness was used to determine the corresponding SAT and PSAT scores using the validity data base from the College Board. For this state assessment, the following empirical definition of college readiness was used:

> XX% probability of a course grade of XX or higher in college level credit bearing courses in the associated subject matter

The 2010 cohort of students was included in the analysis, where their college freshman GPA for ELA and Mathematics and their SAT and PSAT test scores were available. In this dataset, 150 four-year colleges and universities across the country were included, with good representation in terms of location, college size, admission rate as well as a good combination of public and private institutions.

Using PSAT/SAT scores and freshman GPA in ELA and Mathematics, a series of logistic regressions was conducted to determine the scores on the PSAT and SAT that would correspond to the empirical college readiness definition. Once these scores on the SAT and PSAT were determined, the percentage of students above and below the cut scores for the State were identified and incorporated into the standard setting process.

There are a number of limitations to this study, which were explained to the standard setting panel during the process. First, only test scores were taken into account, but not other

relevant factors such as course-taking behaviors. Secondly, only four-year institutions were included in the analysis. Lastly, these percentages were determined using PSAT and SAT, typically for 10th and 11th graders in the state, whereas the assessment we were setting standards on were for grades 3 through 8.

Develop and Review Both Policy and Grade-level Specific PLDs

The policy PLDs were developed by the State Education Department, with input from administrators in the field. The State also consulted with the PARCC policy PLDs (PARCC, 2010), being one of the PARCC states. Four performance levels for reporting were adopted by the State for the following reasons:

- Help schools better target understand students' achievement and provide assistance
- Provide increased opportunities for students, schools and districts to demonstrate growth
- Provide better information across the full range of students' ability level

Below provides an example of the Level 4 (highest achievement level) policy level PLD:

> Students performing at this level **excel** in standards for their grade**.** They demonstrate knowledge, skills, and practices embodied by the State P-12 Common Core Learning Standards for English language arts/literacy that are considered **more than sufficient** for the expectations at this grade.

Using the policy-level PLDs as a foundation, content experts and Department content and assessment experts convened over several months to develop grade- and subject-specific PLDs. These specific PLDs further articulated the knowledge, skills and practices that students performing at a given level should know and be able to do in each content area at each grade level. These grade- and subject-specific PLDs are intended to:

- Provide information to local educators in developing curricular and instructional materials
- Delineate the knowledge, skills and practices that a student should possess at each of the performance level across content areas and grades
- Use as the foundation for the standard setting process
- Provide guidance for future item development for the assessment

The following guidelines were adopted during the development process of the specific PLDs:

- Think big picture!! PLDs are not exhaustive definitions.
- Always think about *observable evidence.*
- Think about what characteristics of one level distinguish it from the adjacent level.
- Use only observable verbs and avoid unobservable verbs
- Describe what students *can* do, not what they *cannot* do
- Avoid adverbs of frequency to discriminate between PLDs (e.g., usually, sometimes, generally) as the assessment is a one-time event.

Below we provide an example of the specific PLDs in grade 3 ELA for all four performance levels:

Anchor standard: students read closely to determine what the text says explicitly and to make logical inferences from it; cite specific textual evidence when writing or speaking to support conclusions drawn from the text.

Level 4: demonstrate an in-depth understanding of a literary text by asking insightful questions and by answering questions, referring explicitly to the text as the basis for the answers.

Level 3: demonstrate a thorough understanding of a literary text by asking and answering questions, referring to the text as the basis for the answers.

Level 2: demonstrate a basic understanding of a literary text by asking factual questions and by answering questions, making inconsistent attempt to refer to the text as the basis for the answers.

Level 1: demonstrate a limited or inaccurate understanding of a literary text by asking and answering questions without referring to the text as a basis for answers.

Convene Standard-Setting Committees

The standard setting event occurred in the summer of 2013. It lasted for five days, with the first four days focusing on the performance standards for specific grades, and the fifth day focusing on the vertical articulation event.

Committee Composition

As noted earlier, the standard-setting process took into account not only the assessed curriculum and content, but also policy considerations and college readiness. As such, the process incorporated input from educators, content experts, policy experts and other stakeholders from the higher education and business communities to help encourage a collaborative effort between a blend of people with content and policy backgrounds. The panelists were also representative of the state geographically and in terms of race/ethnicity. The following provides the composition of the standard setting panels:

- Educators from K-12 and higher education
- Educator administrators, such as principals, curriculum specialists, and superintendents

There were about 25 panelists in each of the standard setting committees. Table leaders were selected to help facilitate group discussions at each table level, so as to ensure all perspectives were included in the standard setting process. Table leaders, along with administrators and policy holders, also stayed the fifth day for the vertical articulation.

Meeting Format and Sequence

Because of the content overlap across grades and the desire to have vertical articulation across grades, the following committees were assembled:

- Committee 1, ELA Grades 3 through 5
- Committee 2, ELA Grades 6 through 8
- Committee 3, Mathematics Grades 3 through 5
- Committee 4, Mathematics Grades 6 through 8

Within each committee, the process started with the performance standards recommendation at the highest grade level, followed by the middle grade, and finally the lowest grade in the assigned grand span.

Introductory General Session

The purpose of the introductory general session was to welcome the standard-setting committees; give background information about legislative requirements, the Common Core aligned assessment program, and the standard setting; and provide the panelists with their charge as the standard-setting committee. A single general session took place for both ELA and mathematics for all grades. The following topics were discussed during the general session:

- Welcome from the State Education Department
- Introduction to the Common Core aligned assessment program
- Background information and the reform agenda

- Panelist charges
- Guiding principles for setting performance standards
- General performance labels and policy definitions
- Overview of standard-setting process
- Overview of standard-setting methodology

Subject-Specific Breakout

After the introductory general session, the panelists broke into their subject/grade-specific committees for the remainder of the meeting. Within each committee, panelists were divided into five table groups. Each table group consisted of both educators and administrators so that there was a blend of education and policy expertise. Such composition at the table level allowed both content and policy perspectives to be considered in the discussions. Table leaders were designated to facilitate the discussions and assist in the meeting logistics (e.g., the collection of judgment forms) at each table.

Once the panelists were in their subject- and grade-specific committees and table groups, information specific to the assessment for which they are setting standards was presented. Each committee member then got the opportunity to review the Ordered Item Book (OIB). The goal of the review was for each committee member to get a feel for the types of items, content, and depth of knowledge on the assessment(s). The test forms that the committee members reviewed were the items administered in the spring.

Threshold Descriptions

The committee members were asked to discuss *key conceptualizations* within their table groups. Through these discussions, committee members who are educators could share their experiences and knowledge from the classroom and on the assessed curricula, while the policy

experts could describe their perspectives on the key considerations and expectations for the program. The goal of this discussion was for committee members to come out of the discussions with some common understanding and expectations for the cut scores that they would be recommending. This is one of the key elements during the standard setting process and a lot of time was allocated for this activity.

To help facilitate the discussions, the panelists were first asked to review the policy, general PLDs, and grade-specific PLDs for the assessment. They then were asked to discuss the following key conceptualization questions at each table:

- What describes a threshold student for each performance category?
- What are the key things that differentiate students from one performance category to the next?
- Which performance level do you believe likely represents the widest range of student ability given the PLDs?
- How different is a student at the very top of the Level 2 from a student at the bottom of the Level 3?

Specific content standards were assigned to each table for the panelists to come up with one or two solid threshold descriptions per standard for each performance level. Once each table finished their discussion, an overall group discussion followed where each table presented their threshold descriptions and discussed those with the rest of the committee.

Discussion on Expectations

Empirical benchmark data were presented during this section. Bar charts, along with descriptions of the analysis, were presented to the panelists. The facilitator engaged panelists in discussions around:

- Are panelists surprised to see the percentage of students based on the college readiness empirical definition?
- How do they expect fifth graders or eighth graders behave given what we see with 10^{th} and 11^{th} graders (PSAT and SAT data)?
- Should Level 3, which is the proficiency level, be linked to college readiness?

Panelists were asked to consider the following when articulating performance expectations for the students:

- Their review of the items and the related skills, knowledge and practices these items are measuring
- Their discussions of the empirical definition and related impact
- What percentage of students would you expect to be in each performance category?
- What, if any, consistency is expected between the impact data from the previous testing program and the new program?
- What type of consistency in pass rates is expected across grades within a content area?
- What, if any, consistency is expected between the assessment data and national assessment data?
- Given that the proficiency level is an indication of students being on track for college readiness, and given the rigor of the common core state standards, please provide your expectation on percentage of students reach proficiency by answering the questions on the rating form

This essentially was a modified version of the Hofstee method (Hofstee, 1983). These data were collected prior to the panelists starting the formal recommendation process. These data were presented, along with the impact data and external benchmark data, later in the process.

Standard-Setting Training

The standard setting committee panelists received training on the *bookmark*, or item-mapping*, procedure* (Lewis, Mitzel, Green & Patz, 1999), which they would use to recommend the cut scores for each assessment. The OIB was one of the primary tools they being used in this process.

The actual OIB consisted of items from the operational administration. In order to eliminate some observed gaps, some of the OIBs were also augmented with some embedded field test items. The external benchmark data were also mapped into the OIB and such bookmarks were made available to the panelists during the standard setting process.

The following guidelines were provided to the panelists for identifying the bookmark they wanted to recommend for the performance standards:

- Think about the knowledge, skills and practices required to answer the item correctly.
- Think about the knowledge, skills and practices of a threshold student at the given performance level.
- Should a threshold student answer this question correctly? That is, should they have a 2/3 chance or greater of answering it correctly?
- If yes, write Yes on bottom of page. If no – write No on bottom of page.

- Go up through item book until you have found the place that you believe best differentiates the items a threshold student should answer correctly from those they are expected to answer correctly.
- Place a bookmark on the last yes item
- Write down the page number on your rating form

Panelists practiced evaluating items and making cut score recommendation in an abbreviated practice OIB in order to try out the item-mapping procedure. Following the practice, the group discussed any questions or difficulties related to the mechanics of the item-mapping procedure. The response probability of 0.67 was used to construct the OIB. This response probability was chosen in order to maximize the amount of information yielded (Huynh, 2000).

Judgment Rounds

The committee panelists went through three individual rounds of judgments for each of the assessments. The panelists started with the highest grade and recommended performance standards. This was in line with the goal of making the assessment program a comprehensive system with performance standards that are aligned and linked to college readiness.

Within each round, the panelists were asked to consider the items in the OIB starting with the easiest item. Each panelist would make a recommendation for the Level 3 (college readiness level) cut score first, followed by a recommendation for the Level 4 (advanced level) cut and a Level 2 cut. Feedback such as individual and group-level cut score recommendations, results from the external benchmark studies and impact data were provided at the end of each round to help inform the panelists' next round of judgments.

In order to make the process transparent and facilitate the panelists' rating process, IRT scale values were printed on the OIB along with the page numbers. The panelists were instructed

on the scale values so they could understand the difficulty level differences of the items. The scale values were introduced so as to help panelists see how different on the IRT scale the two adjacent item pages could be, and this would help panelists better focus their attention. The panelists were also instructed that the page numbers were not raw scores on the test.

Following each round, after feedback reports were provided, panelists engaged in table-level discussions, focusing on items where the group had the largest differences in rating. The facilitator of the standard setting meeting emphasized that consensus was not required; however, these discussions were to help ensure:

- Panelists had a common understanding of the required knowledge, skills and practices the items were measuring
- Panelists were focusing on the requirement of the threshold students and what the threshold students should know and be able to do
- Panelists were bookmarking the last yes item

Once the three rounds completed for a given grade, the standard setting panel moved onto the next assigned grade for their standard setting meeting, and went through the same process starting with discussions on the threshold descriptions for the given grade.

After the panelists completed recommending the performance standards for all three assigned grades, a fourth round occurred, where panelists looked at their round 3 recommendations for all three grades together and provided a round 4 rating, with or without change from their round 3 results.

Vertical Articulation

The final activity in which the standard-setting committee panelists participated was the vertical articulation (smoothing) process. The purpose of vertical articulation was to review the

impact data associated with the recommended cut scores across all grades to see if the trend of the impact data makes sense given the expectations outlined at each grade, the test-taking population, the rigor associated with the content at the different grades, skills/tasks presented on the various assessments, as well as the external benchmark study results. The committee also reviewed and refined the PLDs as necessary so that there was solid alignment between the final committee cut score recommendations and the specific PLDs.

A subset of the standard-setting panelists (table leaders and administrators) participated in the vertical articulation. During the general review session, facilitators informed the panelists that they were now switching gears and reviewing the impact associated with the content-based recommendations in the policy context. The goal was to provide reasonable cut score recommendations to policymakers that consider both the content-based recommendations in addition to their expectations about how students should perform across grades, given the inferences to be made about the assessment.

Panelists were reminded that this was a new and different task. It was important to conduct the vertical articulation for several reasons:

- Many policymakers would be reviewing these recommendations, and the main thing they would be looking at was impact (as is appropriate, because only a content expert can perform the content tasks). If the results did not look reasonable, the policymakers would likely have an issue with the recommendations.
- Two grade-level groups were working independently. Because each group was made up of different panelists, even with the most consistent directions and articulated PLDs, we may still see some inconsistent trends across grades. We, as a group, must

resolve this as best we can. Otherwise, our recommendations would lose some face validity.

Here were the steps followed in the vertical articulation process:

1. Overview the tasks for the articulation activities.
2. Panelists reviewed all the PLDs associated with all grades within the assessment, as well as the threshold descriptions developed during the standard setting. The panelists were asked to especially focus on the PLDs and threshold descriptions from other grades they had not reviewed yet.
3. As a group, asked the panelists to discuss expectations for impact across the grade levels. Remind panelists to consider this discussion within the framework of the PLDs and threshold descriptions, as well as the content coverage and rigor expectation of the assessment across grades. In addition, the panelists were asked to think about external benchmark study results and what it means to be a test assessing skills associated with college readiness.
4. Reviewed the impact associated with the recommended cuts across all grades in side-by-side charts. Discussed the extent to which the data align with their expectations, especially focusing across grade levels and college readiness, if there were consistencies or inconsistencies observed and their alignment with expectations of students' performance across grades.
5. The committee discussed how and if the impact data met expectations, especially in terms of cross-grade consistency and college readiness. The focus was on PLD framework, expectation of students' performance across grades, content justifications, and external benchmark studies.

6. Panelists made independent recommendations as to raw score cuts they believed should be for each grade given the content recommendations, the PLDs, and what they know about the test-taking population, expectations across grades, and inferences on college and career readiness external benchmark studies.
7. Calculated the median raw score recommendations based on the first round of independent rating from the panelists
8. Presented Round 1 result and discussed results as a group.
9. Panelists came to a consensus as to what they believed the final recommended raw score cuts should be
10. Evaluations of the process

In making its recommendation, the policy review committee was asked to consider some of the same types of policy questions posed to the standard-setting committees:

- What percentage of students would you expect to be in each performance category? How does that relate to the college readiness data based on the external benchmark studies?
- What, if any, consistency is expected between the impact data from the previous testing program and the new program?
- What type of consistency in pass rates is expected between courses within a content area?
- What, if any, consistency is expected between State assessment data and data from national or international assessments?

A total of two rounds occurred for the vertical articulation session, one individual rating round and one consensus round. The final consensus recommendation was the final recommendation from the vertical articulation process.

Standards Approval

The Commissioner of Education was presented with the standard-setting panel's Round 4 recommendations, the vertical articulation panel recommendations, external benchmark study results, and the results from the first three rounds. After discussing with his advisers, the Commissioner made the final decision of adopting the recommendations from the vertical articulation panel. These performance standards were presented to the Board and were also formally approved. These performance standards were adopted and reported for the Common Core-aligned assessment.

References

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson and J. S. Helmick (Eds.) On Educational Testing (109-127). San Francisco: Joseey-Bass

Huynh, H. (2000a). On item mappings and statistical rules for selecting binary items for criterion- referenced interpretation and Bookmark standard settings. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA

McClarty, Way, Porter, Beimers, & Miles (2013). Evidence Based Standard Setting: Establishing a Validity Framework for Cut Scores. *Educational Researcher, 42*(2), 78-88.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

PARCC PLDs (2013). Retrieved from: https://www.parcconline.org/plds

Wiley, A, Wyatt, J., & Camera, W.J. (2011). The Development of a Multidimensional College Readiness Index. Retrieved from: https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2010-3-development-multidimensional-college-readiness-index.pdf