Measuring the Power of Learning.™

Research Memorandum
ETS RM–15-07

Alignment Between Innovative Summative Assessment Prototypes and the Common Core State Standards: An Exploratory Investigation

Richard J. Tannenbaum

Patricia A. Baron

Priya Kannan

September 2015

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Alignment Between Innovative Summative Assessment Prototypes and the Common Core State Standards: An Exploratory Investigation

Richard J. Tannenbaum, Patricia A. Baron, and Priya Kannan
Educational Testing Service, Princeton, New Jersey

September 2015

Corresponding author: Richard J. Tannenbaum, E-mail: rtannenbaum@ets.org

Suggested citation: Tannenbaum, R. J., Baron, P. A., & Kannan, P. (2015). *Alignment between innovative summative assessment prototypes and the Common Core State Standards: An exploratory investigation* (Research Memorandum No. RM-15-07). Princeton, NJ: Educational Testing Service.

Action Editor: James Carlson

Reviewers: Clyde Reese and Randy Bennett

Abstract

In this study we collected teachers' judgments about the alignment between innovative, computer-delivered summative assessment prototypes (Cognitively Based Assessment *of, for,* and *as* Learning [*CBAL*™]) and the Common Core State Standards (CCSS). The assessments included scenario-based tasks (with embedded questions) with content components distributed across assessment occasions. One set of 4 assessments addressed Grade 8 writing, and 1 set of 2 assessments addressed Grade 7 reading. These prototype assessments had been developed before the CCSS were available. The goal of the study was to investigate the extent of convergence between the assessments and the CCSS. Teachers judged (a) the consistency between the knowledge and skills of the assessment questions and those defined by the CCSS (content overlap) and (b) the consistency between the cognitive complexity of the questions and that expressed by the CCSS (level overlap). Results indicated that across the writing tasks, 5 of the 10 CCSS writing standards were addressed. Across the reading tasks, 5 of 9 CCSS literature standards and 7 of 10 informational text standards were addressed.

Key words: alignment, Cognitively Based Assessment *of, for,* and *as* Learning (*CBAL*™), Common Core State Standards, summative assessment

Cognitively Based Assessment *of, for,* and *as* Learning (*CBAL*™) is a research-based initiative that offers schools a unique opportunity to measure what kindergarten through 12th grade (K−12) students know and can do at multiple points during the school year, not just once at the end, which is typical practice. Bennett and Gitomer (2009) noted several benefits of a periodic-assessment model. For one, collecting information at multiple points in time enables teachers to modify instructional practice to assist their current group of students. When information is collected only once, at the end of a year, teachers have no opportunity to react to their current students' learning needs. Further, rather than one assessment carrying the sole weight of a summative decision, the weight is distributed across the multiple assessments and a summative decision becomes the product of the accumulated evidence from these assessments. Multiple assessments within a domain also increase the ability to cover content in greater depth than is feasible within a one-time assessment model. Periodic assessments, as Bennett and Gitomer noted, provide students and teachers the opportunity to recover from any one unrepresentative assessment outcome and likely increase the reliability of any summative inferences.

The conceptualization and construction of CBAL assessments have been informed by competency models (Deane et al., 2008; O'Reilly & Sheehan, 2009). Each model is domain-specific (e.g., writing) and is derived from content standards, learning-sciences research, and if available, learning progressions for that domain (Bennett, 2010). The two sets of prototype assessments, writing (Grade 8) and reading (Grade 7), which are the focus of this research study, were developed before the Common Core State Standards (CCSS). The CCSS represent a standardized expectation of what students at different grade levels should know and be able to do if they are to be considered on track to be career and college ready by the conclusion of their secondary education.

The purpose of this study is to collect evidence from panels of teachers (with no prior involvement in the CBAL initiative) on the perceived alignment between the writing Grade 8 CCSS and CBAL prototype writing Grade 8 assessments and between the reading Grade 7 CCSS and CBAL prototype reading Grade 7 assessments. This study is exploratory, with the goal of identifying the points of convergence and gaps between the assessments and the CCSS. Because the assessments were developed before the release of the CCSS, gaps were anticipated with the expectation that the results of the study would inform future revisions of these assessments.

Two features make this study particularly informative and timely. The first is that the design structure of the prototypes roughly parallels that of the Common Core State Assessments of the two large consortia: Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment. In both cases, the content domains to be covered are measured by assessments delivered at more than one time point. In the case of PARCC, for example, the time points are after 75% and 90% of the instructional year has elapsed. This structure, as noted, should enable the measurement both of the breadth and depth of content domains. The second feature, also shared with the Common Core State Assessments, is that many of the CBAL prototype assessment tasks measure multiple competencies simultaneously. More traditional assessments tend to have items that are intended to measure one competency each, which makes the task of judging alignment less complex and demanding. However, given the emphasis in standards directed at the integration of skills, it is more likely that future assessments will move further away from the one-item-one-competency model.

CBAL Assessments

The assessments include a variety of task types: selected-response, technology-enhanced (e.g., drag-and-drop), short constructed-response, and extended constructed-response. The tasks are computer delivered and responses are in most cases computer scored. As noted, any one task may measure more than one competency. The use of separately administered components (periodic assessments) means that interpretations of alignment need to be considered across the components, as they would for the discrete item and performance components of the Common Core State Assessments. The CBAL assessments are described in brief below.

Writing (Grade 8)

Four assessments measure the construct of writing at the eighth-grade level. The first, ban ads, focuses on writing arguments and includes four tasks composed of 16 questions. The second, mango street, focuses on literary analysis and includes four tasks composed of 13 questions. The third, invasive species, focuses on research and exposition and includes four tasks composed of 32 questions. The fourth, service learning, focuses on argument and exposition and includes four tasks composed of 11 questions.

Reading (Grade 7)

Two assessments measure the construct of reading at the seventh-grade level. One assessment, seasons, is directed at literary reading. There are four tasks composed of 20 questions. The second assessment, wind power, targets the reading of informational text. This assessment has five tasks composed of 20 questions.

Methods

Alignment is a judgment-based process (Eckhout, Plake, Smith, & Larsen, 2007). In the context of student assessment, most often teachers with expertise in the content area and teaching experience at the relevant grade level are convened as a panel; between five and eight educators typically form a panel (N. L. Webb, 2007). Specific approaches to alignment may vary (e.g., Bhola, Impara, & Buckendahl, 2003; A. Porter, McMaken, Hwang, & Yang, 2011; A. C. Porter, 2002; N. L. Webb, 1999, 2007; N. M. Webb, Herman, & Webb, 2007). However, what is fundamental to each approach is a comparison between assessment questions and a set of standards that delineate content expectations and expectations about cognitive complexity or depth of understanding (Herman, Webb, & Zuniga, 2007). Alignment information is considered a key source of validity evidence, as it directly addresses score interpretations and uses (Bhola et al., 2003; Davis-Becker & Buckendahl, 2013; Martone & Sireci, 2009).

In this study, we modified the basic alignment procedures suggested by N. L. Webb (1999, 2007). Our approach focuses on panelists making judgments of (a) the degree to which the knowledge and/or skills required to respond to a question overlap with the kinds of knowledge and/or skills defined by the CCSS (content overlap) and (b) the degree to which the cognitive complexity of a question is *below*, *at*, or *above* that expressed by the CCSS (level overlap). (The cognitive complexity of a question is a joint function of the task stimuli and question.) Each of the writing task questions was evaluated against 10 Grade 8 writing standards; each of the reading task questions was evaluated against nine Grade 7 literature standards (one standard, argument, is not applicable to literature) and 10 Grade 7 informational text standards.

Panelists

One panel of nine teachers participated in the alignment for writing. A separate panel of eight teachers participated in the alignment for reading. None of the teachers had been previously involved in either the writing or the reading assessment development. The teachers were

recruited from states within commuting distance to the meeting site (Connecticut, New Jersey, New York, and Pennsylvania). All teachers had grade-level relevant English language arts teaching experience (current or in the past 3 years). During the recruitment process, teachers were asked to report their level of familiarity with the CCSS using the following scale: 1 (*not familiar*), 2 (*less than somewhat familiar*), 3 (*somewhat familiar*), 4 (*more than somewhat familiar*), 5 (*familiar*). The majority of writing teachers (7 of 9) reported being familiar with the CCSS. The majority of reading teachers (6 of 8) reported that they were more than somewhat familiar with the CCSS.

Procedures

The alignment for writing and reading occurred concurrently. One researcher facilitated the writing panel and another researcher the reading panel. Each panel met for 2 days and followed the same alignment procedures. The common alignment process is described next.

Prior to the alignment study. The teachers received a copy of the CCSS relevant to their grade-level assessments and were asked to review the standards. They were encouraged to highlight key terms or phrases or to otherwise make notes on the standards. The intent of the exercise was to reinforce their understanding of the standards. The teachers were also required to take the assessments for which they would be conducting the alignment. The assessments were made available online. The teachers took the assessments under untimed conditions and had the opportunity to take the assessments multiple times. The goal was for the teachers to engage with the assessments to become familiar with the content and the complexity of the tasks and questions.

During the alignment study. The rationale for conducting the alignment study was explained to the teachers. (The purpose of the study had been shared with the teachers during their recruitment for potential participation in the study.) The first two activities of the study focused on the teachers discussing the CCSS and the assessments. The teachers shared their interpretations of what the standards emphasized and valued and what was expected of students to meet the standards. The goal was for the teachers to recognize how their colleagues understood the standards. The discussion of the assessments focused on the major knowledge and/or skills being measured across the assessments, and the extent to which the content they covered was consistent with the knowledge and/or skills they expected of their students. The goals were to help the teachers recognize, at a molar level, what was (and was not) being

measured by the assessments and to consider connections to the learning goals they have for their students. Together with the prestudy assignments, the discussion of the standards and the assessments helped ensure that the teachers had a reasonable understanding of the two main sources of alignment.

The teachers received training on the alignment process, which included making independent practice alignment judgments and discussing their judgment rationales for one task/question for one of the assessments. The process of alignment is question-based and occurs for one assessment at a time. Each question within a task is evaluated in relation to each of the specific standards for that domain and grade. For example, there are 16 questions for the ban ads Grade 8 writing assessment and there are 10 Grade 8 writing standards. A teacher independently evaluates Question 1 alignment with each of the 10 writing standards, then the teacher independently evaluates Question 2 against each of the 10 standards, and so on through Question 16. In essence, the teachers used a matrix of questions by standards. A question could be judged to be aligned to more than one standard.

The teachers first judged the degree to which the knowledge and/or skills required to respond to the question overlapped with the kinds of knowledge and/or skills defined by the standards (content overlap). The teachers were to indicate if the content overlap was 1 (*low*), 2 (*medium*), or 3 (*high*). If no overlap was perceived between a question and a standard, the teacher left the response cell blank. The second alignment judgment focused on the degree to which the cognitive complexity of a question was 1 (*below*), 2 (*at*), or 3 (*above*) that expressed by the standards (level overlap). This second judgment was only rendered if, first, the teacher perceived at least a low level of content overlap. In other words, if the content of a question did not align with the content of a standard, no judgment of level overlap was needed.

Two rounds of alignment judgments occurred for each assessment. The first round was independent. Once the Round 1 judgments for an assessment were completed, the results were computed and presented to the teachers for feedback and discussion. The feedback was a completed ratings matrix of questions by standards. For each question, the number of teachers assigning content overlap ratings of *low*, *medium*, or *high* and the number of teachers assigning level overlap ratings of *below*, *at*, or *above* were revealed for each standard. Figure 1 is an example of the response feedback for Question 1 (of Task 1A) for the ban ads writing assessment. The first column under each standard displays the content overlap ratings across the

nine teachers; the second column displays the level overlap ratings. Recall, if a teacher did not perceive any content overlap, that teacher left the response cells blank both for C (content overlap) and L (level overlap). As illustrated, four teachers judged there to be a *low* level of content overlap between Question 1 and Standard 1; one judged there to be a *medium* level of content overlap, and two judged there to be a *high* level. For that same standard, four teachers (not necessarily the same four that rated the question *low* for content overlap) judged the question to be *below* the expectation of the standard. The cognitive demand of the question was less than that defined by Standard 1. Two teachers, however, judged the question to be *at* the same level as the standard, and one teacher perceived that the question was *above* (exceeded) the cognitive expectation of the standard. None of the teachers judged there to be any alignment between Question 1 and Standard 3. Once this type of feedback was displayed, the teachers shared the rationales for their judgments. The teachers were then given an opportunity to make a second round of judgments.

Standards

	1		2		3		4		5		6		7		8		9		10	
	C	L	C	L	C	L	C	L	C	L	C	L	C	L	C	L	C	L	C	L
Task 1A																				
Q1																				
✓	4	4	3	4			2	1	1	1	2	2	1	1	1	4	2	3	1	1
✓✓	1	2	2	2			2	5			1	1			4	1	5	5	1	1
✓✓✓	2	1	1				2										1			

Figure 1. Example of response feedback. C = content overlap; L = level overlap; ✓ = *low* and *below*; ✓✓ = *medium* and *at*; ✓✓✓= *high* and *above*.

In this study, Round 1 judgments are considered preliminary, because the teachers had not yet had the opportunity to engage in discussions to clarify their own ratings or to seek clarification from other teachers on the panel. While the Round 1 ratings are independent, the goal of this type of alignment study is to obtain the most reasoned set of judgments, which is facilitated by the between-round discussions. Round 2 ratings, therefore, are considered the operational set of alignment decisions. Once the Round 2 judgments were completed for one assessment, the process was repeated for the other assessments. At the conclusion of each study, each teacher completed a form evaluating the quality of their alignment-study experience.

Results

The same analyses were conducted for the writing-alignment judgments and the reading-alignment judgments. We have focused the analyses on the Round 2 ratings, as they are the operational ones. We tallied the number of teachers providing *low*, *medium*, or *high* content-overlap ratings and the number providing *below*, *at*, or *above* level-overlap ratings for each question by standard (see Figure 1 for an example for one question). These frequency counts are the fundamental data points. We identified instances where a majority of teachers on a panel (at least five) rated either *medium* or *high* content overlap and/or rated either *at* or *above* level overlap. N. L. Webb (2007) suggested that positive evidence of overlap exists if at least six questions are judged to address the same standard. The rationale for this metric builds from the work of Subkoviak (1988),[1] indicating that six questions measuring the same standard offer a reasonable lower limit of reliability. We applied this rule of thumb to interpret our alignment results, although we recognize that others have proposed different criteria (e.g., Norman & Buckendahl, 2008). We present the results first for writing and then for reading.

Writing

Table 1 presents the instances where a majority of teachers (at least five) rated either *medium or high* content overlap and/or rated *at* or *above* level overlap between questions and standards. (The labels for each standard are our short-hand descriptors. Each panelist worked with the actual CCSS. The same applies to the reading standards.) Standards 1, 2, 4, and 9 were addressed by more than six questions both in terms of content overlap and level overlap, with a large number of questions (68 for content and 49 for level) covering Standard 9 (textual evidence). Standard 8 (use of sources) was addressed by 33 questions in terms of content, but by

only five questions in terms of level. This disparity was due to only two questions from the invasive species assessment meeting the level of the standard; 29 questions from this assessment met the content-overlap criterion. Standard 3 (narrative) was not aligned with any questions. Standards 5 (writing process) and 6 (technology) were only addressed by one question; Standards 7 (research) and 10 (range of writing) also were underrepresented. The numbers in Table 1 represent an aggregation across questions within each assessment. (The individual task/question-to-standard results were provided to assessment developers to inform potential modifications to the assessments.)

Table 1. Writing Standards: Number of Questions Judged by a Majority of Panelists to Have *Medium* or *High* Content Overlap and/or *At* or *Above* Level Overlap With Standards

Standards	Argument–Ban ads 16 Qs		Literary analysis–Mango street 13 Qs		Research/exposition–Invasive species 32 Qs		Argument/exposition–Service learning 11 Qs		Total # of Qs meeting standard (max. possible = 72)	
	C	L	C	L	C	L	C	L	C	L
1. Argument	3	4	2	2	1	1	2	2	8	9
2. Exposition	4	4	3	2	3	3	2	2	12	11
3. Narrative										
4. Writing quality	4	4	3	3	3	3	2	2	12	12
5. Writing process					1	1			1	1
6. Technology					1	1			1	1
7. Research		1			2	3			2	4
8. Use of sources			2	2	29	2	2	1	33	5
9. Textual evidence	15	13	13	8	31	23	9	5	68	49
10. Range of writing			2	1	1	1	1	1	4	3

Note. Q = question; C = content; L = level.

Reading: Literature Standards

Table 2 presents the instances where a majority of teachers (at least five) rated either *medium* or *high* content overlap and/or rated either *at* or *above* level overlap between questions and standards. Standard 1 (textual evidence) was addressed by 25 questions for content overlap and 24 questions for level overlap. Standard 10 (range of literary texts) was addressed by 19 questions both for content overlap and level overlap; as may be expected, this result was due solely to the seasons assessment, which focuses on literary text. Standards 2 (main ideas), 3 (interactions), and 4 (vocabulary) were addressed by between 8 and 12 questions. Standards 7 (text v. other modes) and 9 (comparative analysis) were not aligned with any questions.

Standards 5 (text structure) and 6 (point of view) also were underrepresented. The seasons assessment was aligned to the largest number of standards, which is not surprising given that it was intended to address literary text. Interestingly, the wind power assessment, which was intended to focus on informational text, contributed the most hits for Standard 1 (textual evidence).

Table 2. Reading Literature Standards: Number of Questions Judged by a Majority of Panelists to Have *Medium* or *High* Content Overlap and/or *At* or *Above* Level Overlap With Standards

Standards	Literary text– Seasons 20 Qs		Informational text– Wind power 20 Qs		Total # of Qs meeting standard (max. possible = 40)	
	C	L	C	L	C	L
1. Textual evidence	10	10	15	14	25	24
2. Main ideas	5	5	6	4	11	9
3. Interactions	8	9			8	9
4. Vocabulary	9	11		1	9	12
5. Text structure	1	3			1	3
6. Point of view	1	2			1	2
7. Text v. other modes						
9. Comparative analysis						
10. Range of literary texts	19	19			19	19

Note. Standard 8 (argument) does not apply to literature (http://www.corestandards.org/ELA-Literacy/RL/7) and was not included in this table. Q = question; C = content; L = level.

Reading: Informational Text Standards

Table 3 presents the instances where a majority of teachers (at least five) rated either *medium* or *high* content overlap and/or rated either *at* or *above* level overlap between questions and standards. Standard 1 (textual evidence) and Standard 10 (range of nonfiction texts) were each addressed by more than 20 questions. Standards 3 (interactions) and 4 (vocabulary) were each addressed by between 12 and 20 questions. Standard 2 (main ideas) was addressed by nine questions for content and 10 for level. Standard 8 (argument) was addressed by six questions each for content and level. Standard 7 (text v. other modes) was not aligned with any questions. Standards 6 (point of view) and 9 (comparative analysis) also were underrepresented.

Table 3. Reading Informational Text Standards: Number of Questions Judged by a Majority of Panelists to Have *Medium* or *High* Content Overlap and/or *At* or *Above* Level Overlap With Standards

Standards	Literary text– Seasons 20 Qs		Informational text– Wind power 20 Qs		Total # of Qs meeting standard (max. possible = 40)	
	C	L	C	L	C	L
1. Textual evidence	10	9	15	14	25	23
2. Main ideas	2	4	7	6	9	10
3. Interactions	7	8	5	6	12	14
4. Vocabulary	12	10	8	4	20	14
5. Text structure	2	4	4	4	6	8
6. Point of view			3	4	3	4
7. Text v. other modes						
8. Argument			6	6	6	6
9. Comparative analysis		1	3	3	3	4
10. Range of nonfiction texts			20	20	20	20

Note. Q = question; C = content; L = level.

Conclusion and Discussion

This study was designed to evaluate the perceived alignment between the CBAL assessments for writing Grade 8 and reading Grade 7 against the corresponding CCSS. This study was exploratory in nature, as the assessments had been constructed before the development of the CCSS. Two panels of teachers, one for writing and one for reading, participated in the alignment evaluation process.

Across the writing assessments (Table 1), five of the 10 writing standards were addressed by a relatively large number of questions: Standards 1 (argument), 2 (exposition), 4 (writing quality), 8 (uses of sources), and 9 (textual evidence). This result was especially true for Standard 9, for which 68 questions were judged to be aligned at either a *medium* or *high* content level, and 49 questions were judged to be *at* or *above* the complexity defined by the standard. Of course, one may consider this large number of "hits" from the perspective of construct underrepresentation. That is, given limits on the time available for assessment, the emphasis being devoted to this one standard can be viewed as displacing the measurement of other standards.

With respect to such displacement, significant gaps between the assessments and the CCSS were noted. The writing results indicate that, for example, Standards 3 (narrative), 5

(writing process), and 6 (technology) were not well represented by assessment tasks. No questions were considered to be aligned with Standard 3, and Standards 5 and 6 received only one hit each for content and level. If the intended objective going forward is to measure the full complement of CCSS Grade 8 writing standards, assessment developers would need to focus primary attention on building tasks that measure these standards. Additionally, Standards 7 (research—"Conducting short research projects . . .") and 10 (range of writing— "Write routinely over extended time frames . . .") were weakly addressed by the assessment tasks. Fewer than six hits were evidenced for each standard; again, these would be areas where task revision or development would be warranted.

The reading assessments addressed five of nine literature standards (excluding Standard 8, argument, which is not relevant to literature) and seven of the 10 informational text standards. Standard 7 (text v. other modes) in both sets of standards addresses comparing and contrasting a written text to one or more alternate forms of that text (e.g., video, audio); none of the reading tasks was judged to measure this competency. Literature Standards 5 (text structure), 6 (point of view), and 9 (comparative analysis) were similarly underrepresented, especially Standard 9, which had no hits. This standard, comparative analysis, involves comparing and contrasting a fictional portrayal with a historical account. The informational text standards were better represented. The only standard in addition to Standard 7 (text v. other modes) not sufficiently addressed (fewer than six hits) was Standard 9 (comparative analysis). For this set of reading standards, comparative analysis concerns analyzing how multiple authors present different stances on the same topic.

Alignment focuses on the interaction between questions and standards. The emphasis is on the content of a question, the knowledge and skills it measures, and how that measurement converges with the knowledge and skills defining the standards. The writing and reading assessments included a variety of question types, with the majority being a variation of selected-response questions. The choice of question type is informed by the content or skills intended to be measured, what is needed to elicit the targeted content or skills, but also by testing-time and scoring constraints. The choice of question type seems to have impacted the writing-alignment judgments more so than was observed for the reading assessments. The question-specific alignment judgments (reported to the assessment developers) clarify that the writing constructed-response questions (either short or extended) were associated with the largest number of

standards. In contrast, the writing selected-response questions were associated with far fewer standards. These outcomes are not surprising, given how the writing standards are defined. The writing standards place a value on active writing—write argument, write informative/explanatory text, write narratives, produce clear and coherent writing. The constructed-response questions, by their nature, require test takers to demonstrate aspects of their active writing competence. The selected-response questions within each writing assessment do address important elements of writing (e.g., the critical thinking component skills associated with analyzing and critiquing arguments), which the panel of teachers recognized and acknowledged in their discussions; however, the teachers found it difficult to reconcile the active writing espoused by the standards with the process of choosing a response, even a technology-enhanced, selected-response. The response format was not a factor for the reading alignment judgments.

Koretz and Hamilton (2006) raised awareness of the challenges of interpreting alignment outcomes when, as they indicated, results may vary due to the complexity of the alignment model implemented, the number of standards considered, and the particular rating scales that may be used. They further noted the absence of well-formed guidance on how to handle innovative tasks that, like the scenario-based tasks within CBAL, may be designed to tap into more than one content standard and perhaps more than one level of cognitive complexity, or that may predispose an assessment toward a particular segment of the standards (e.g., informational reading). The complexities of alignment evaluations in such contexts notwithstanding, this study identified areas in which CBAL reading and writing assessments would need to be supplemented in order to fully cover the CCSS, which could be fulfilled with more traditional items specifically designed to pull in standards beyond those addressed by the scenario-based tasks. But the goal of full coverage of standards places increased demands on testing time, as shown by the Common Core State Assessments, which may take several hours to complete.

The judgmental nature of alignment studies makes critical the training of the educators to complete the alignment task, as designed. Training includes not only explanation of the purpose of the alignment and description of the steps to follow, but also an opportunity for educators to practice making the alignment judgments and to provide rationales for their judgments. It is through this practice and discussion that misconceptions are revealed, which can then be rectified. This process of training-practice-discussion is similar to what occurs in standard-setting studies, which are also dependent on professional judgment (Tannenbaum & Katz, 2013).

Further, if the intent of the alignment is to obtain the most reasoned estimate of the perceived overlap between assessments and standards, then it seems appropriate to include more than one round of judgments with feedback and discussion between rounds. Here, again, we see a parallel to standard-setting practices, where at least two rounds of judgments are considered best practice (Plake, 2008). The first-round outcomes, although reflective of independent judgments, are part of the educators' learning process of the judgment task. The second round may, therefore, reflect a more informed set of results because the educators have received feedback on their first-round judgments, have had a chance to discuss these judgments with one another, and have had the opportunity to consider taking those discussions into account in their second round. Training and multiple rounds of judgment are but two implementation features common both to standard setting and alignment. Davis-Becker and Buckendahl (2013) identified others and suggested that comparable evaluative criteria may be applied to each activity such as indicators of procedural appropriateness and decision consensus. In alignment, the former, for example, refers to methods that consider both content and cognitive complexity matches, and the latter to the inclusion of rules regarding panelist agreement; both of these quality criteria were accounted for in the current study. Further, although our study implemented a variation of N. L. Webb's (2007) alignment approach, the included design features are applicable to other judgment-based alignment practices.

References

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement, 8,* 70–91.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43−61). New York, NY: Springer.

Bhola, D., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21–29.

Davis-Becker, S. L., & Buckendahl, C. W. (2013). A proposed framework for evaluating alignment studies. *Educational Measurement*: *Issues and Practice, 32*(1), 23–33.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (Research Report No. RR-08-55). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02141.x

Eckhout, T. J., Plake, B. S., Smith, D. L., & Larsen, A. (2007). Aligning a state's alternative standards to regular core content standards in reading and mathematics: A case study. *Applied Measurement in Education, 20,* 79–100.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education, 20,* 101–126.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Education Research, 79,* 1332–1361.

Norman, R. L., & Buckendahl, C. W. (2008). Determining sufficient measurement opportunities when using multiple cut scores. *Educational Measurement: Issues and Practice, 27*(1), 37–45.

O'Reilly, T., & Sheehan, K. M. (2009). *Cognitively Based Assessment of, for, and as Learning: A framework for assessing reading competency* (Research Report No. RR-09-26). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02183.x

Plake, B. S. (2008). Standard setters: Stand up and take a stand. *Educational Measurement: Issues and Practice, 27*(1), 3–9.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher, 40,* 103–116.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31,* 3–14.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability for mastery tests. *Journal of Educational Measurement, 25,* 47–55.

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Madison: Wisconsin Center for Education Research, University of Wisconsin.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20,* 7–25.

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice, 26*(2), 17–29.

Notes

[1] This criterion was initially proposed to guide decisions regarding subscore reporting. If the goal is only to represent the breadth of the standards and not to report a score on each individual one, fewer items per standard would arguably be justifiable.