



Evaluating the Alignment of Mathematics, Reading, and Science Field Test Items to the South Dakota Content Standards

Final Report

Brett P. Foley, Ph.D.
Susan L. Davis, Ph.D.
Chad W. Buckendahl, Ph.D.
Deirdre A. Lupper, M.Ed.

January 6, 2012

Executive Summary

A project was undertaken to inform the South Dakota Department of Education (SDDOE) about how Math, Reading, and Science field test items aligned with the South Dakota content standards and the Common Core State Standards. South Dakota educators and content specialists with experience at the relevant grade levels met in Sioux Falls, SD on October 25-26, 2011 to provide judgments that were used in the alignment analysis.

In this process, alignment judgments for the items were made in terms of content and cognitive complexity. This report documents the number of items judged to align to the content standards by both grade level and content area and provides item-level content and cognitive complexity judgments.

Across grade levels and content areas, the summary results of the judgments made by the educators generally supported evidence of alignment for most items. In total, the panels reviewed 804 items across 17 subject/grade levels - the panelists identified alignment of 620 of these items to the South Dakota content standards and identified alignment of 521 of these items to the Common Core content standards (most Reading items aligned to both sets of standards). Across grades and subjects, between 0% and 44% of the panels' judgments disagreed on cognitive complexity with the judgments from earlier bias and content review panels; between 0% and 45% of the panels' judgments disagreed with the earlier panels for content. Therefore, most items could be included in South Dakota's item bank to be considered for inclusion on future forms of the state's assessments (assuming all other psychometric characteristics of the items are acceptable). Alignment judgment results are discussed in the body of the report. Panelists also anonymously completed an evaluation of the alignment process. Results of the evaluation suggested that the panelists understood the activity and had confidence in their content fit and cognitive complexity judgments.

Results of this project should be useful to SDDOE to inform assessment form development, to revise items that do not match their specifications, and to improve future item development processes to increase the proportion of acceptable items. This report also serves as independent validity evidence for the state regarding the alignment of their items with the state content standards. Because this project was limited to reviewing item level alignment information, inferences about the alignment of existing intact forms of these assessments should not be made; however, these results are expected to facilitate evaluation of existing and future assessment forms by providing the independent, item-level judgments needed for a complete form-alignment study.

Evaluating the Alignment of Mathematics, Reading, and Science Field Test Items to the South Dakota Content Standards

Final Report

As part of the *No Child Left Behind* (NCLB) legislation, states are required to provide information about the technical quality of their content standards and assessments. One specific requirement of NCLB's peer review process is for states to provide evidence of an independent alignment of their assessments to their state-adopted content standards.

Many states have used variations of a method described by Webb (1997) to evaluate the evidence of alignment between assessment items and content standards. Other methodologies have also evaluated the intersections among instructional content, curriculum, and assessments (Porter, 2002). Broadly stated, these methods could be interpreted as extending content validity research to include other dimensions that may be relevant for supporting inferences about students' scores.

Frisbie (2003) also proposed a method for conducting alignment studies that explicitly evaluates the cognitive complexity (CC) of content standards and assessment items, and the content fit of assessment items to the content standards. From a data collection perspective, this method could be characterized as moderately complex (Bhola, Impara, & Buckendahl, 2003) because it focuses on both content and CC of both the content standards and the assessment items. Given that this study focused on evaluating the alignment of field test items, a modified version of Frisbie's approach was utilized.

In addition to providing independent alignment judgments, panelists were asked to consider the alignment judgments provided by an earlier group of subject matter experts who were empaneled to provide a content and bias review of the new test content. Both the initial independent judgments and the final judgments after considering those of the earlier content and bias review panel are provided in this report.

The results of this project should be useful to SDDOE to inform assessment development and to serve as independent validity evidence for the state regarding the alignment of their assessment items with the Math, Reading, and Science content standards. This feedback can be used to inform assessment form development, revise items that do not match their specifications, and improve future item development processes to increase the proportion of acceptable items. Because this project was limited to reviewing item-level alignment information, inferences about the alignment of existing intact forms of these assessments should not be made; however, these results are expected to facilitate evaluation of existing and future assessment forms by providing the independent, item-level judgments needed for a complete form-alignment study.

Procedures

SDDOE recruited a total of 44 South Dakota educators from public and private schools as well as regional support centers to participate in this alignment study as panelists. This group represented the primary geographic regions of the state proportional to the different sizes of schools within the state. In addition, panelists were familiar with the population of students within their respective content area and grade level. The panelists were teachers, teacher leaders, subject coaches, department chairs, and/or content specialists.

The panelists had extensive experience in education. A summary of the education, current positions, and average years of experience by alignment panel is shown in Table 1. Additional information on the recruiting and demographic characteristics of the panelists is on file with SDDOE.

Table 1. Experience and Background by Panel

	Highest Degree Earned			Current Position		Average Years of Experience
	N	Bachelors	Masters	Teachers	Teacher Leader/ Subject Coach/ Department Chair/ Specialist	
Math Grades 3/4/5	6	5	1	6		18.8
Math Grades 6/7	7	5	2	7		14.9
Math Grades 8/11	6	3	3	6		16.7
Reading Grades 3/4/5	5	2	3	3	2	18.0
Reading Grades 6/7	7	4	3	6	1	11.3
Reading Grades 8/11	7	4	3	7		11.4
Science Grade 5/8/11	6	2	4	6		9.0

At the start of the meeting, panelists received a packet of materials that included an agenda, demographic questionnaire, alignment rating instructions, rating forms, copies of the content standards documents, a listing of the CC of the content standards, and an evaluation form that was completed at the end of the process. A confidentiality and non-disclosure form was distributed and signed by participants prior to receiving any proprietary materials. The meeting began with a welcome from Gay Pickner, SDDOE’s Assessment Director. Susan Davis-Becker from Alpine Testing Solutions (Alpine) provided an overview of the alignment study purpose and conducted training for the panelists on the alignment methodology and procedures.

Prior to the workshop, each panelist had been pre-assigned to a content- and grade-span panel. Panels tasked with three sets of field test items included: Reading 3/4/5, Math 3/4/5, and Science

5/8/11¹. Panels were tasked with two sets of field test items including: Reading 6/7, 8/11, and Math 6/7, 8/11. After the orientation, the panelists were divided into three groups based on content area. The first group included the three Math panels (3/4/5, 6/7, 8/11). The second group included the three Reading panels (3/4/5, 6/7, 8/11). The third group included the Science panel (5/8/11).

Frisbie's (2003) method was utilized in the sense that item-level judgments were gathered on the alignment of items based on both content and cognitive complexity. SDDOE has identified the need to assess students against both the South Dakota Content Standards (SD Standards, see www.doe.sd.gov/contentstandards) and the Common Core Standards (<http://www.corestandards.org/>). For Reading, SDDOE identified overlap between the SD Standards and the Common Core standards. Therefore, SDDOE created the new Reading items to align to both sets of content standards. For Mathematics, SDDOE recognized differences between the SD Standards and the Common Core standards and therefore created a unique item set to align to each framework. Because Common Core standards for Science do not yet exist, these items were only intended to align to the SD standards.

Additionally, SDDOE wanted to evaluate the match of the new items to the content standards in terms of cognitive complexity. The SD Standards were created using Bloom's Taxonomy (Bloom, 1956). Therefore, the items that were written to align to the SD Standards were evaluated for cognitive complexity using Bloom's framework. In contrast, the Common Core standards were identified using Webb's Depth of Knowledge (Webb's DOK, 1997) framework for cognitive complexity. Therefore, the items that were written to align to the Common Core Standards were evaluated using Webb's DOK. Descriptions of both CC frameworks can be found in Appendix A. Operationally, this led to all Reading items being compared to both sets of standards and both CC frameworks as these items were written to address both sets of standards. For Science, all items were compared to the SD Standards and Bloom's taxonomy only. For Math, those items written to address the SD Standards (about half of the items) were compared to the SD Standards and Bloom's taxonomy whereas the items written to address the Common Core Standards (the other half of the items) were compared to the Common Core Standards and Webb's DOK.

Within their subject/grade level groups, a panel's first activity was to review the cognitive complexity framework that they would use to evaluate the items and to review the cognitive complexity levels of the content standards (all groups began with SD Standards and Bloom's taxonomy).

The panelists' second task was to review each of the field test items and make their alignment ratings. In conducting this task, panelists were not provided information about the intended alignment or any other prior alignment judgments for these items. Panelists first rated the cognitive complexity of the item and recorded their judgments on a custom-designed Alpine rating form. Panelists then identified the standard, if any, to which the item aligned in terms of content. They were asked to evaluate the fit of the items within standard using the following dichotomous judgment:

¹ The Science alignment study was initially designed as inclusive of three panels. Small samples for each panel led to the combination of the three for the operational activities.

Complete or Partial Fit:

The content of the item fits completely (or substantially) within the standard.

No Fit/Slight Fit

The content of the item does not fit within any standard (or only a small part of the item fits within the standard).

If panelists judged the item as completely or partially fitting the standard, they recorded the associated content standard number on the rating form. If the panelists only found a slight fit with the content standards or found no match, they were instructed to mark an "X". Panelists made both judgments (cognitive complexity and content) independently and then discussed them as a panel to achieve group consensus, which was recorded by a panel leader. The panel was encouraged to discuss items for which they did not agree, but were also reminded that consensus did not require unanimous agreement, only a simple majority among members of the panel.²

For a third and final judgment, panels revisited their ratings for items where their consensus judgments (in terms of cognitive complexity, content, or both) disagreed with the earlier proposed alignment classifications from the bias and content review panels. The process proceeded as follows. After the alignment panels conducted their group discussions, Alpine staff compared their consensus judgments with the earlier classifications that resulted from bias and content review, where these bias and content review judgments were made by different panels of South Dakota educators facilitated by Pearson. Any disagreements identified (cognitive complexity, content, or both) were noted by the Alpine facilitator and shared with the alignment panel. The alignment panels were asked to consider the classifications from the bias and content review panels and could either modify or keep their original group consensus ratings. Pearson staff members were available throughout the alignment meetings to answer panelists' questions regarding the earlier item classifications. The results presented in this report are based on the alignment panels' final consensus decisions after considering the earlier bias and content panels' judgments.

Each panel completed the alignment process of field test items being compared to the SD Standards and Bloom's taxonomy on the first day, identifying content fit between the items and the content standards, coming to consensus on each of these judgments within their panel, and reviewing results from earlier panels when there were disagreements. At the end of the first day, the Science panel had completed their judgments for the field test items. The remaining panels (Reading and Math) returned on the second day to complete the process for the items using the Common Core State Standards and Webb's DOK. When each panel had completed their alignment work, they completed an evaluation of the process. Following the workshop, Alpine compiled the consensus alignment results which are presented below.

² This definition was applied throughout the study whenever group "consensus" was used to determine the panels' recommendations.

Results

Alignment

In this section we present the summary results, by grade and subject, in terms of the overall content alignment of the items to the standards. Because these are groups of field test items, rather than intact forms, only item-level results are meaningful. Tables 2-4 indicate the number of items, by content and grade level, that matched at least one of the content standards. Based on the results in Tables 2-4, the vast majority of items were found to align to at least one standard. In total, the panels reviewed 804 items across 17 subject/grade levels - the panelists identified alignment of 620 of these items to the SD standards and identified alignment of 521 of these items to the Common Core standards (most Reading items aligned to both sets of standards).

Table 2. Alignment Results Summary – Math Content

Standards	Grade	Number of items	Content Alignment	
			Matched at least one standard	No Match
SD	3	20	20	-
	4	20	20	-
	5	20	20	-
	6	21	21	-
	7	20	20	-
	8	20	20	-
	11	19	19	-
Common Core	3	25	25	-
	4	25	25	-
	5	24	24	-
	6	30	30	-
	7	28	28	-
	8	29	29	-
	11	21	21	-

Table 3. Alignment Results Summary – Reading Content

Standards	Grade	Number of items	Content Alignment	
			Matched at least one standard	No Match
SD	3	49	48	1
	4	51	51	-
	5	52	52	-
	6	46	45	1
	7	49	49	-
	8	50	50	-
	11	57	57	-
Common Core	3	49	44	5
	4	51	51	-
	5	52	50	2
	6	46	44	2
	7	49	46	3
	8	50	49	1
	11	57	55	2

Table 4. Alignment Results Summary – Science Content

Standards	Grade	Number of items	Content Alignment	
			Matched at least one standard	No Match
SD	5	37	37	-
	8	45	45	-
	11	46	46	-

In Tables 5-7, results from the alignment study are compared to the alignment judgments from the earlier bias and content review panels. After initial group consensus judgments, between 0% and 62% of the items at each subject/grade/standard were judged to be at a different cognitive complexity level than was indicated by the bias and content review panels. With respect to content alignment, between 0% and 69% of the items at each subject/grade were aligned to a different standard than was indicated by the bias and content panels. These items are labeled as "initial" disagreements in Tables 5-7. The alignment panels reviewed the discrepancies between their results and the results from the bias and content panels and adjusted their ratings where they felt it was appropriate. Following this review process, the incidence of disagreements between panels was reduced to between 0% and 44% of the items for cognitive complexity and between 0% and 45% of the items for content. These items are labeled as "final" disagreements in Tables 5-7. Some alignment disagreements are expected because alignment is an inherently subjective process. For example, disagreements may arise from ambiguity in content standards or from varying views on the steps students will employ to arrive at an answer for an item. Similar disagreements between panelists and panels have been reported in the alignment literature (see, for example, Webb, Herman, and Webb, 2007; Wyse and Viger, 2011).

Table 5. Comparison of Math Alignment results with earlier Bias/Content Review results

Standards/ CC Framework	Grade	Number of items	Disagreements			
			CC		Standard	
			Initial	Final	Initial	Final
SD/Bloom's	3	20	50%	20%	10%	5%
	4	20	45%	25%	15%	15%
	5	20	15%	-	5%	5%
	6	21	43%	10%	14%	-
	7	20	55%	20%	-	-
	8	20	40%	10%	-	-
	11	19	53%	26%	-	-
Common Core/Webb's DOK	3	25	36%	-	4%	4%
	4	25	20%	16%	4%	-
	5	24	17%	17%	-	-
	6	30	40%	7%	10%	3%
	7	28	14%	-	36%	-
	8	29	52%	17%	17%	14%
	11	21	-	-	33%	10%

Table 6. Comparison of Reading Alignment results with earlier Bias/Content Review results

Standards/ CC Framework	Grade	Number of items	Disagreements			
			CC*		Standard	
			Initial	Final	Initial	Final
SD/Bloom's	3	49	NA	NA	49%	41%
	4	51	NA	NA	53%	37%
	5	52	NA	NA	38%	25%
	6	46	NA	NA	35%	17%
	7	49	NA	NA	31%	18%
	8	50	NA	NA	46%	16%
	11	57	NA	NA	53%	16%
Common Core/Webb's DOK	3	49	43%	18%	45%	20%
	4	51	61%	31%	55%	45%
	5	52	35%	10%	69%	33%
	6	46	46%	17%	54%	53%
	7	49	45%	18%	29%	25%
	8	50	62%	24%	52%	24%
	11	57	44%	23%	30%	16%

* Bias/Content Review results using Bloom's were not available for the reading items

Table 7. Comparison of Science Alignment results with earlier Bias/Content Review results

Standards/ CC Framework	Grade	Number of items	Disagreements			
			CC		Standard	
			Initial	Final	Initial	Final
SD/Bloom's	5	37	51%	41%	5%	3%
	8	45	51%	44%	-	-
	11	46	46%	39%	26%	9%

The full results of the alignment study are included in Appendix B. The included results files contain detailed item-level alignment judgments for both content and cognitive complexity. This feedback can be used to revise items that do not match their specifications and perhaps amend the process by which they were developed to improve the proportion of acceptable items in the future.

Evaluation

Following the alignment study, panelists completed an evaluation of the process. This information is used to judge the success of the process and can be used to improve future alignment studies. Table 8 shows the median response for each evaluation question, by subject

area. Overall, the panelists' indicated the training was successful and the right amount of time was devoted to training. In the second part of the evaluation, the panelists indicated they were confident in their judgments and the right amount of time was devoted to making judgments. In the final part of the evaluation the panelists indicated they felt the process was successful and well organized. Specific feedback was provided by the Science and Reading (elementary grade-span) panels on the amount of time provided for the work that had to be accomplished. Although both groups finished within the time allocated for the meeting, their comments will be considered when planning future meetings.

Table 8. Alignment study evaluation results

Evaluation Question	Median Response		
	Math	Reading	Science
Number of Completed Evaluations	19	19	6
Success of Each Training Component <i>6 = Very Successful, 1=Very Unsuccessful</i>			
Orientation	5	5	5
Overview of Standards	5	5	5
Discussion of the Rating Scales	5	5	5
Practice Rating Activity	5.5	5	4.5
Overall Training	6	5	4.5
Time Allocated to Training <i>3=Too much 2=Right amount, 1=Too little</i>	2	2	1.5
Alignment Judgments – SD Standards and Bloom’s taxonomy			
Confidence in Judgments <i>4=Very Confident to 1 = Not at all Confident</i>			
Content Fit of Items to Standards	4	4	4
Cognitive Complexity of Items	4	3	3
Time Allocated to Making Judgments <i>4=More than enough to 1 = More time needed</i>	3	3	2.5
Alignment Judgments – Common Core Standards and Webb’s DOK			
Confidence in Judgments <i>4=Very Confident to 1 = Not at all Confident</i>			
Content Fit of Items to Standards	3	3	NA
Cognitive Complexity of Items	3	3	NA

Evaluation Question	Median Response		
	Math	Reading	Science
Time Allocated to Making Judgments <i>4=More than enough to 1 = More time needed</i>	3	3	NA
Overall Success of the Study <i>4=Very Successful to 1 = Very Unsuccessful</i>	4	3	3
Overall Organization of the Alignment Study <i>4=Very Successful to 1 = Very Unsuccessful</i>	4	4	3

In addition to their evaluation ratings, panelists were encouraged to provide comments on the process. The comments provided are listed below by subject area and topic (experience, process improvement):

Math

- *Experience*
 - I absolutely loved this! I have no comments for improvement. It was perfect.
 - Extremely helpful! I'm so glad I came. Very insightful and makes me less fearful of the new standards.
 - This was awesome! I have heard about the biased content review and data review and would love to come.
 - I learned so much!
 - Very stressful.
 - Great experience. Thanks for the opportunity.
 - Very good opportunity
 - Great PD!
 - Great!!
 - Very organized
- *Process Improvement*
 - When lunch is provided, we don't need an hour lunch
 - Really don't need a whole hour for lunch when it is provided on site. 1/2 hour and then back to work would allow earlier finish. :)
 - More snacks

Reading

- *Experience*
 - Things were well organized. Leaders were definitely open to ideas! Good job.
 - Very beneficial & useful workgroup. Helped me better understand and analyze common core & SD standards in relation to test questions
 - We had to do 3-5 reading and we had to do 3 days of work in two days. We were here hours after the other groups and felt very stressed to get done.
- *Process Improvement*

- Reading grades 3-5: Even though we were extremely diligent, we could not do justice to the elementary grades because the task was overwhelming. I suggest that another independent group also align, and compare/contrast to our alignment. A much better scenario would be: option 1 3 days for reading 3-5; option 2, grades 3-4 2 days, grades 5-6 2 days, grades 7-8 2 days, grade 11 1 day
- Add another day for work or split the elementary section
- Organize the workshop so 1 group does not do 3 grade levels. For example, set it up as follows, grades 3/4 2 days, grades 5/6 2 days, 7/8 2 days, 11 1 day
- The reading group of 3-5 really needed another day to do justice to our work. We had more questions than everyone else and should have had more time allotted. It was hard to focus when everyone else got to leave--when we are getting the same pay! You should the groups differently grades 3/4, grades 5/6, grades 7/8 2 days; grade 11 one day
- Please consider grouping questions together with the reading assign they go with
- Provide copies of "unpacked" standards. All questions on a selection together in review book
- Have the state/common "unpacked" standards available/square tables for easier access of all documents
- Better climate control. Square table would facilitate better group interaction.

Science

- *Process Improvement*
 - It would have been helpful to have more time with specific grade levels, rather than try and do all three grade levels. By the time we got to 11th grade my brain was fried
 - Science group had a lot of work to do. We were fried at the end. Other resources to help with Bloom's may have been useful.
 - I didn't feel comfortable doing HS standards. Get more HS staff in
 - If science is going to align all 3 grade levels, 2 days need to be provided. 11th grade received much less concentration.
 - A more consistent, clear guideline for cognitive complexity is needed. Also, should be split 5th/8th/11th teachers, not a mix.

Conclusions and Recommendations

The purpose of this project was to identify the extent to which the Mathematics, Reading, and Science field test items aligned with the South Dakota content standards in Mathematics and Reading at grades 3-8 and 11, and Science at grades 5, 8, and 11. The results suggest that most items are aligned with the content standards at each of the grade levels.

These results should be useful to SDDOE to inform assessment form development, to revise items that do not match their specifications, and to improve future item development processes to increase the proportion of acceptable items. In addition, this information can be used to facilitate future test form development/alignment analysis. Completion of this study by an independent organization provides another source of validity evidence for the state regarding the alignment of their items with the state content standards.

References

- Bhola, D. S., 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.
- Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives, the classification of educational goals – Handbook I: Cognitive Domain*. New York, NY: McKay.
- Frisbie, D. A. (2003). Checking the alignment of an assessment tool and a set of content standards. Iowa Technical Adequacy Project (ITAP). Iowa City, IA: University of Iowa.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31(7), 3-14.
- Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education (NISE Research Monograph No. 6). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.
- Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, 26, 17-29.
- Wyse, A. E. & Viger, S. G. (2011). How item writers understand depth of knowledge. *Educational Assessment*, 16, 185-206.

APPENDIX A – Cognitive Complexity Frameworks

The files below contain the Webb's DOK and Bloom's Taxonomy descriptors provided to panelists as guidance.



Webb's DOK
Descriptors



Bloom's Taxonomy
Descriptors

APPENDIX B –Alignment Results

The embedded Excel documents contain the full item-level results of the alignment study. Within the file, the results for each subject/grade are represented on a separate worksheet. Both cognitive complexity and aligned standard are included. Items were allowed to be matched to more than one standard. An "X" in the standard column indicates the panelists determined that the item did not fit any of the content standards for that grade level.



Mathematics



Reading



Science