

**REPORT ON THE ALIGNMENT ANALYSIS OF
CURRICULUM STANDARDS AND *WESTEST* IN
MATHEMATICS, READING LANGUAGE ARTS, AND
SCIENCE FOR GRADES 3, 4, 5, 6, 7, 8, AND 10,
WEST VIRGINIA**

Norman L. Webb

April 30, 2004

Table of Contents

Executive Summary

Introduction..... 1

Mathematics..... 1

Reading Language Arts..... 8

Science 13

The Alignment Process 20

Conclusions..... 20

References..... 22

Appendices

Executive Summary

Three groups of eight content experts, four from West Virginia and four from other states, analyzed the agreement between the West Virginia academic standards and the WESTEST. One group analyzed the agreement in each of three content areas—mathematics, reading language arts, and science. The analyses were done in two stages, once in July, 2003, and once in November, 2003. During the intervening period, state staff worked with the testing vendor to improve the standards and assessment alignment based on the information provided in the first analysis. The assessment system, including the academic standards and assessments for grades 3, 4, 5, 6, 7, 8, and 10 and the three content areas, showed coherence in that students' progress could be tracked from one grade to the next and greater sophistication in the content area was required as students progressed from grade to grade. In judging the alignment among the academic standards and assessments, breadth, depth, and emphasis were analyzed. Cognitive complexity was compared using four depth-of-knowledge levels. The reviewers found strong alignment in mathematics. In reading language arts, the alignment of the two of the three standards appropriately measured with on-demand assessments, and the assessments overall, showed good and acceptable alignment. In science, the assessments adequately measured the two most significant standards and parts of the other four standards. Thus, the alignment was good for the two standards, but could be improved by either combining two or more of four standards, or by increasing the number of items measuring these four standards. The alignment in science also could be improved by including a few assessment items with higher depth-of-knowledge levels.

REPORT ON THE ALIGNMENT ANALYSIS OF CURRICULUM STANDARDS AND WESTEST IN MATHEMATICS, READING LANGUAGE ARTS, AND SCIENCE FOR GRADES 3, 4, 5, 6, 7, 8, AND 10 IN WEST VIRGINIA

Norman L. Webb
April 30, 2004

Introduction

This is a report on two alignment studies that analyzed the agreement between the expectations for student learning as described in the West Virginia Academic Standards in reading/language arts, mathematics, and science with the state WESTEST in grades 3 through 8 and 10. The alignment analyses were done in two institutes four months apart to provide time for adjustments to the assessments in order to improve the match between the assessments and the academic standards and to confirm that improvement. The analyses addressed and reported on the categories of content represented in both the standards and the assessments, the depth-of-knowledge in each, the breadth or range of the content in the standards measured by the assessments, and the emphasis given on the assessments to one or more content areas under a standard. As such, the alignment analyses produced results on the comprehensiveness of the assessments, content and performance match, the degree of emphasis in the standards represented by the assessments, the cognitive depth and breadth of the assessments in relation to the standards, and consistency with the achievement standards. The findings of the alignment study were reported to be clear to all members of the school community.

Overall, the analyses of alignment between the West Virginia academic standards and WESTEST assessments found that:

1. In mathematics, there was strong alignment.
2. In reading/language arts, the alignment with the two of the three standards targeted by the assessment was good and acceptable.
3. In science, the assessments represented the desired emphasis among standards as indicated in the science blueprint, but that the alignment in other ways could be improved.

The results of the alignment analyses are discussed for each of the three content areas.

Mathematics

At the first alignment analysis, the eight reviewers reached consensus on the depth-of-knowledge level for each objective under each of the five standards for each grade level prior to coding the items. The results from their deliberation are presented in Exhibit 1. Across the grades, reviewers rated 75% to 90% of the objectives at depth-of-knowledge (DOK) levels of 1 (Recall) and 2 (Skills and Concepts). Reviewers' analyses indicated some progression across grades, finding that the latter grades had a higher

percentage of objectives rated at Level 3 (Strategic Thinking). In grades 7, 8, and 10, more than 20% of the objectives were rated at Level 3.

In the increased percentage of objectives with high depth-of-knowledge levels and in the graph (Exhibit 2), it is evident that the assessment system in mathematics has attributes that provide for a coherent assessment system. There is a clear progression from the lower to higher grades in the depth-of-knowledge levels of the expectations for student learning. The higher grades include a greater proportion of content to be achieved at Level 2 (Skills and Concepts) and Level 3 (Strategic Thinking). The proportion of items on the assessments from grade 3 through grade 8 and grade 10 also shift appropriately among the five standards. The number of items on the test forms across the grades represents greater emphasis in Number and Operations (Standard I) in the lower grades and a higher emphasis on Algebra (Standard II) in the higher grades, with an even emphasis across the grades on Geometry (Standard III), Measurement (Standard IV), and Data Analysis/Probability (Standard V) (Exhibit 3). The distribution of items by standard across the grades corresponds to the design of the assessments as indicated in the test blueprints developed by the state.

Exhibit 1

Percent of Objectives by Depth-of-Knowledge (DOK) Levels for Each Grade, West Virginia Alignment Analysis for Mathematics

Grade	Number of Objs	DOK Level	# of objs by Level	% w/in std by Level
Grade 3	44	1	20	45
		2	18	40
		3	5	11
		4	1	2
Grade 4	44	1	23	52
		2	16	36
		3	4	9
		4	1	2
Grade 5	37	1	14	37
		2	20	54
		3	3	8
Grade 6	34	1	14	41
		2	18	52
		3	2	5
Grade 7	36	1	12	33
		2	16	44
		3	8	22
Grade 8	32	1	4	12
		2	20	62
		3	8	25
Grade 10	25	1	2	8
		2	17	68
		3	6	24

Exhibit 2
Mathematics DOK Levels for Objectives by Grade

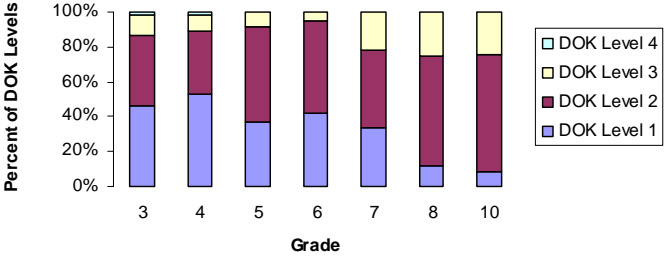


Exhibit 3
Percent of Hits by Standard for Each Grade

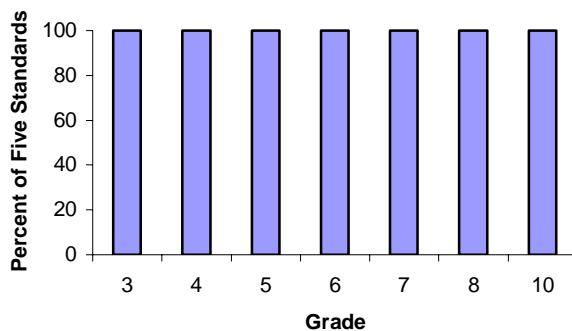
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Number/Operations	42	40	37	26	23	18	18
Algebra	13	15	16	26	26	30	34
Geometry	18	19	17	19	15	18	19
Measurement	16	13	16	14	16	14	16
Data Analysis/Prob	11	13	14	15	20	20	13

Each form of the assessment for grades 3 through 8 had 52 items, 47 multiple-choice items and five open-response items. The grade 10 assessment had 49 items, 44 multiple-choice items and five open-response items. Each open-response item was worth three points. The total points for each assessment for grades 3 through 8 was 62 points and for grade 10 was 63 points. In comparing the computations of the assessment and curriculum standards in the alignment analysis, all of the items were regarded as equally weighted.

The alignment in mathematics between the assessments and the standards for all seven grades is good. For all grades, each form had a sufficient number of assessment items that corresponded to each standard to satisfy the Categorical Concurrence criterion of six or more items per standard (Exhibit 4). An acceptable level on the Categorical Concurrence criterion indicated that the assessment has a minimal number of items measuring content for a standard in order to make some judgment about a student’s mastery related to the standard. Thus, the WESTEST in mathematics is a comprehensive assessment because it measures content corresponding to each of the five mathematics standards and it has a sufficient number of items for each standard as a basis for making judgments about students’ knowledge of content related to each standard.

Exhibit 4

Percent of Five Mathematics Standards with An Acceptable Number of Items by Grade

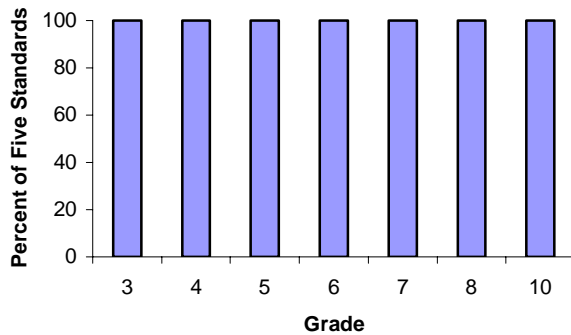


For each grade, the items were adequately distributed among the objectives under each standard by measuring content of at least 50% of the objectives under a standard to satisfy the Range-of-Knowledge Correspondence criterion (Exhibit 5). This indicated that

the breadth of items on the assessment represented a similar breadth of content as described in the five mathematics standards and was consistent with the range of content as represented in the standards. Thus, the content sampled on the assessment does measure students' knowledge of content on all of the standards and, within each standard, a range of content that produces results that reflect the meaning of the content standards.

Exhibit 5

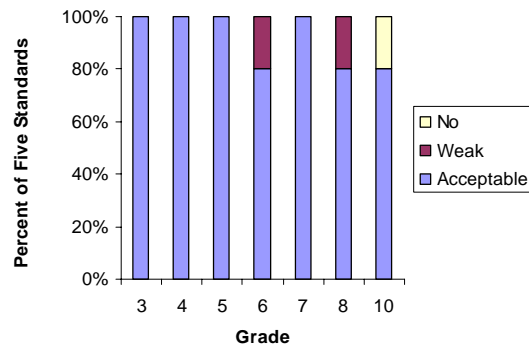
Percent of Mathematics Standards with Acceptable Range-of-Knowledge Correspondence by Grade



As indicated in Exhibit 3, the proportion of assessment items measuring content related to a standard varied by grade in a pattern that corresponds to the design of the assessments. Within a standard, the alignment analysis assumed that the emphasis given to each of the objectives under a standard should be equal. A balance index was computed to depict the degree to which one or two objectives were over-emphasized by an excessive number of items compared to other objectives under a standard. In general, across the grades and standards, the items did not overemphasize one or two objectives compared to the other objectives. However, for one standard in each of three grades—Number and Operations (Standard I) for grades 6 and 10 and Measurement (Standard IV) for grade 8—one of the possible objectives was overemphasized on the assessments compared to the other objectives under the standard with corresponding items (Exhibit 6). Thus, across the seven grades analyzed, more than 90% of the standards had assessment items equally distributed among the underlying objectives, which was considered an appropriate emphasis between the assessment and standards. The distribution of items among the standards and among the objectives within each standard indicated that the assessments reflected the same degree of emphasis as the standards.

Exhibit 6

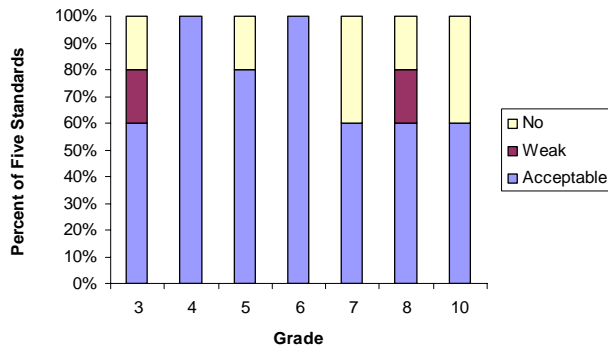
Percent of Mathematics Standards by Acceptable Level of Balance of Representation by Grade



The depth-of-knowledge measured by the assessment items compared favorably to the expected depth-of-knowledge for 75% of the standards of the 35 analyzed (five standards for each of seven grades) (Exhibit 7). For five of the seven grades, too few of the assessment items measuring content related to Data Analysis/Probability (Standard V) were at or above the depth-of-knowledge level of the corresponding objective. Most of the objectives under the Data Analysis/Probability standards expected students to do work at Level 2 (Skills and Concepts), or Level 3 (Strategic Thinking). However, a number of the Data Analysis/Probability assessment items were judged to be at Level 1 (Recall). For the three higher grades (grades 7, 8, and 10), an insufficient number of assessment items measuring content related to Number and Operations (Standard I) were at or above the depth-of-knowledge level of the corresponding objective. In the higher grades, an increasing percentage of objectives under Standard I were judged to be at a DOK level of 2 or 3, whereas the decreasing number of assessment items measuring content related to this standard were judged to be Level 1. The alignment analysis identified two specific areas where the agreement between the depth measured by the assessments and the depth expected by the standards could be improved, Standard V (Data Analysis/Probability) and Standard I (Number and Operations). This shortfall in the assessments was judged insufficient to deem the standards and assessments as inadequately aligned. Rather, with the assessments measuring content at an appropriate depth-of-knowledge level for three-quarters of the standards (27 out of 35 standards), the assessments have a reasonable depth compared to the standards.

Exhibit 7

Percent of Mathematics Standards by Acceptable Level of Depth-of-Knowledge Consistency by Grade



Overall, the analysis of the West Virginia operational tests and curriculum standards in mathematics for seven grades indicated that the alignment was reasonable and good. At each grade level, there were a sufficient number of items to judge attainment of a standard, and these items were adequately distributed among the objectives to measure a range of the required knowledge under each standard. For nearly all of the standards across the seven grades, the items matching objectives were distributed fairly evenly. The depth-of-knowledge (DOK) levels of a small number of items, mainly those relating to Data Analysis/Probability standards, were lower than the level of corresponding objectives. Thus, the Depth-of-Knowledge Consistency criterion was not fully met for the Data Analysis/Probability standard for five of the seven grades. The number of alignment issues was low and could be altered by replacing at most four items on any of the test forms. It was concluded that the WESTEST assessments in mathematics for grades 3, 4, 5, 6, 7, 8, and 10 are aligned with the West Virginia curriculum standards.

Reading Language Arts

For Reading Language Arts, eight reviewers reviewed the objectives under the three language arts standards. They reached consensus on the depth-of-knowledge level for each objectives. These levels are presented in Exhibit 8. In grades 3, 4, and 5, reviewers rated from 70% to 80% of the objectives at a DOK level 1 or 2, mainly Recall/Recognition and Conceptual/Procedural understanding. In the later grades, the proportion of objectives at DOK Level 3, Inference and Analysis, increased (Exhibit 9). The percentage of objectives rated at a DOK level 3 or 4 increased from 17% at grade 3 to 50% or higher at grades 8 and 10, which is characteristic of a coherent system that shows progression.

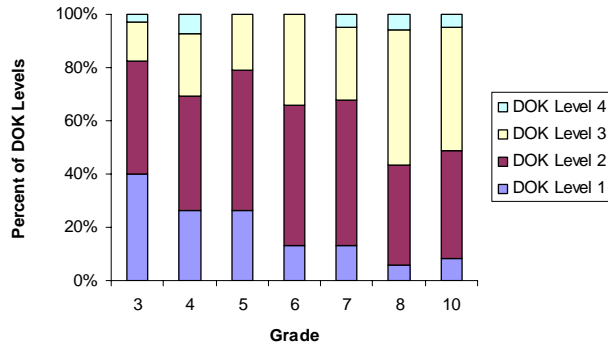
Exhibit 8

Percent of Objectives by Depth-of-Knowledge (DOK) Levels for Each Grade, West Virginia Alignment Analysis for Reading Language Arts

Grade	Number of Objectives	DOK Level	# of Objs by Level	% w/in Standard by Level
Grade 3	28	1	11	39
		2	12	42
		3	4	14
		4	1	3
Grade 4	26	1	7	26
		2	11	42
		3	6	23
		4	2	7
Grade 5	38	1	10	26
		2	20	52
		3	8	21
Grade 6	38	1	5	13
		2	20	52
		3	13	34
Grade 7	37	1	5	13
		2	19	52
		3	10	27
		4	2	5
Grade 8	32	1	2	6
		2	12	37
		3	16	50
		4	2	6
Grade 10	37	1	3	8
		2	15	40
		3	17	45
		4	2	5

Exhibit 9

Reading Language Arts DOK Levels for Objectives by Grade



The assessment was designed to measure student knowledge in reading and language arts would be reasonable with an on-demand assessment. The assessment was not designed to assess students’ facility in listening, speaking, and viewing. Items included on the assessment measured what students know and can go in the area of two of the three standards, Reading (Standard I) and Writing (Standard II), and not for the Listening, Speaking, and Viewing (Standard II). Over the seven grades, the reviewers rated about 60% to 70% of the items corresponding to reading and the remaining items corresponding to writing objectives (Exhibit 10). This indicates that a consistent level of importance was given to reading and writing across the grades. However, as noted above, an increasing proportion of the objectives across the grades required students to engage in inference and analysis. The alignment analysis provided evidence that the West Virginia Reading Language Arts standards and assessment show coherence across grade levels and that the assessment met the design attributes, indicated in the blueprint, of the distribution of items among two of the three Reading Language Arts standards.

Exhibit 10

Percent of Hits by Standard for Each Grade and the State’s Blueprint Designed Range

		Grade 3	Grade 4	Grade 5 ¹	Grade 6 ¹	Grade 7 ¹	Grade 8	Grade 10
Reading	Study	68%	67%	57%	59%	64%	66%	60%
	Blueprint	65-70%	65-70%	60-70%	65-70%	65-70%	65-70%	60%
Writing	Study	32%	33%	43%	41%	36%	34%	40%
	Blueprint	30-35%	30-35%	30-40%	30-35%	30-35%	30-35%	40%
List/Speak/View	Study	0	0	0	0	0	0	0
	Blueprint	0	0	0	0	0	0	0

¹ Considering the five 2-point items for each grade, the proportion of items by standard varies some: Grade 5 Reading (59%) and Writing (41%); Grade 6 Reading (60%) and Writing (40%); Grade 7 Reading (65%) and Writing (35%).

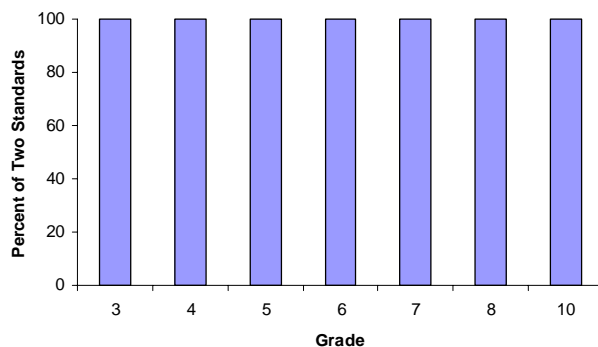
The number of items on each form of the assessment analyzed varied some in the total number of items by grade—grade 3 (70), grade 4 (75), grade 5 (76), grade 6 (75),

grade 7 (75), grade 8 (75), and grade 10 (80). At each grade the assessment included five open-response items, each worth two points. Thus, the total points possible ranged from 75 (grade 3) to 85 (grade 10). The computations of the comparison of the assessment and curriculum standards in the alignment analysis considered all of the items as weighed equally.

The alignment of the West Virginia WESTEST and curriculum standards in reading language arts for seven grades was found to be reasonable in general, given constraints on the assessment for the Reading and Writing standards. The test form analyzed included a large proportion of items that measured comprehension skills and strategies at all grade levels. The overemphasis in this area was by design and did not compromise the relevance of the assessment in making appropriate inferences about students' attainment of the Reading standard. Also, by design, no items on the assessment were found to measure content related to the Listening, Speaking, and Viewing standard. Multiple forms of the test at each grade were used, although not all forms measured the same objectives. In the one form at each grade that was analyzed, there was a sufficient number of items measuring content related to each of the two standards to meet the Categorical Concurrence criterion at an acceptable level, six or more items per standard (Exhibit 11). Thus, the assessments had an adequate number of the items on which to base judgment regarding students' attainment of the standards.

Exhibit 11

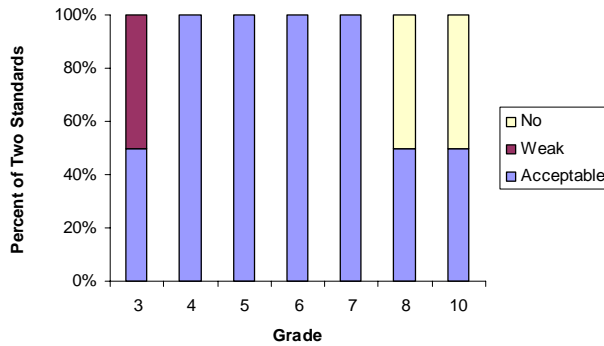
Percent of Two Reading Language Arts Standards with An Acceptable Number of Items by Grade



For grades 4 through 7 in particular, the items on the assessments were adequately distributed among objectives to meet the minimum requirements for the Range-of-Knowledge criterion (Exhibit 12). For the grade 3 Writing standard, this criterion was only weakly met. The assessment items on the form analyzed for grades 8 and 10 did not measure content of an adequate proportion of objectives under the Writing standard, 50% or more of the objectives. However, it was noted that other writing objectives were measured on other forms that were not part of the analysis.

Exhibit 12

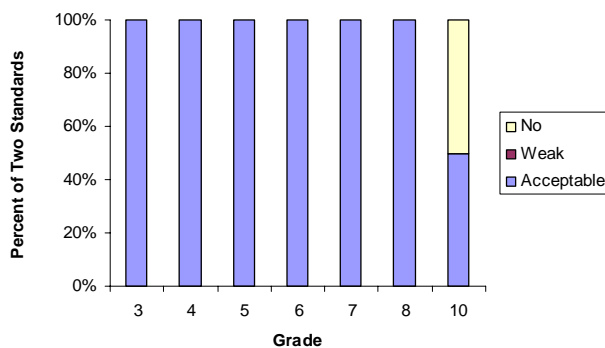
Percent of Two Reading Language Arts Standards with An Acceptable Range-of-Knowledge Correspondence by Grade



In the analysis of the reading language arts standards, there was good agreement between the depth-of-knowledge levels of the objectives under the two standards and the DOK levels of the corresponding items. Of the 14 analyses, two per grade, the assessments and two standards met an acceptable level on 93% of the analyses (Exhibit 13). Only for grade 10 did the analysis indicate that the depth-of-knowledge levels of the assessment items did not sufficiently match the depth-of-knowledge levels of corresponding objectives under the Reading standard—50% of the items with a DOK level at or above the DOK level of the corresponding objectives. As indicated in Exhibit 9, the DOK levels of objectives across the grades increased in complexity by including a

Exhibit 13

Percent of Reading Language Arts Standards by Acceptable Level of Depth-of-Knowledge Consistency by Grade

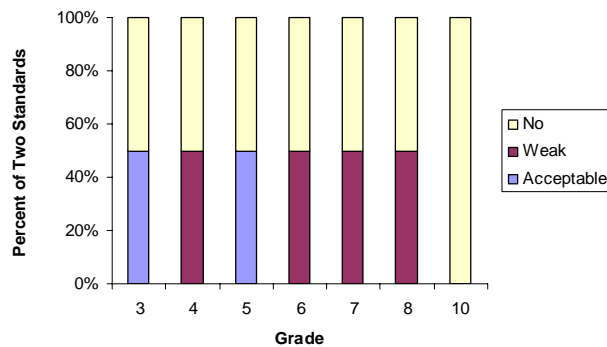


higher proportion of objectives at DOK Level 3, using inferences and strategic thinking. For grade 10 writing, the assessment reflected this increase in complexity, but not for grade 10 reading. For grade 10 reading, most of the items had a DOK level of 2, Skills and Concepts. However, at the other grades, the progression in complexity on the assessments corresponded to the intended complexity as expressed by the standards.

The degree of emphasis on the assessment given to specific objectives varied considerably within each grade, resulting in a low proportion of acceptable levels on the Balance-of-Representation criterion (Exhibit 14). For all seven grades, a large number of items corresponded only to one or two of the more than 10 objectives under the Reading standard. This resulted in an unacceptable level on the Balance-of-Representation criterion, which assumed an even distribution of items among the objectives tested. These over-emphasized objectives, in general, related to the application of comprehension skills or strategies. Because the Reading standard and assessment achieved an acceptable level on the other three criteria, a skewed distribution of items can be regarded as appropriate if this is desired by West Virginia. The assessment and Writing standard were found to have an acceptable level of alignment on the Balance-of-Representation criterion for grades 3 and 5. On the other grades, the Balance-of-Representation criterion was only weakly met for four of the grades and not met for grade 10. Where the Reading standard is concerned, a skewed distribution of items is not considered a fatal flaw in alignment as long as the other criteria are met. This was not the case for grades 8 and 10. At these grades, a large proportion of the items corresponding to the Writing standard measured one or two objectives, while more than 50% of the Writing objectives had no corresponding items. This was seen as a shortfall and was identified as an area for improvement, even if some of the objectives were covered on other forms.

Exhibit 14

Percent of Reading Language Arts Standards by Acceptable Level of Balance of Representation by Grade



Overall, the alignment between the West Virginia Reading Language Arts Standards and one form of the WESTEST was, in general, reasonable, given the constraints on the assessment. The on-demand assessment does not include any items measuring content related to one of the three Reading Language Arts Standards—Listening/Speaking/Viewing. Content from this standard can be better assessed by other means than by a paper-and-pencil test that is machine-scored. There are an adequate number of items on the assessments at all seven grade levels to measure students’ attainment of the other two standards, Reading and Writing. For all of the grades, except

for the grade 10 Reading standard, the items are at a depth-of-knowledge level that adequately matches the depth-of-knowledge levels of the corresponding objectives. At four of the grades—grades 4, 5, 6, and 7—an adequate proportion of the objectives had at least one corresponding item for the Range-of-Knowledge Correspondence criterion to be met for both the Reading and the Writing standards. The distribution of items could be improved for the other three grades, but because other objectives were assessed on other forms, the lack of attainment of the range criterion for these three grades was not considered critical. By design, some areas under the Reading and Writing standards were emphasized more on the assessment than other areas. Because an acceptable alignment was generally achieved on the other criteria across the grades, the imbalance was not viewed as significant.

Science

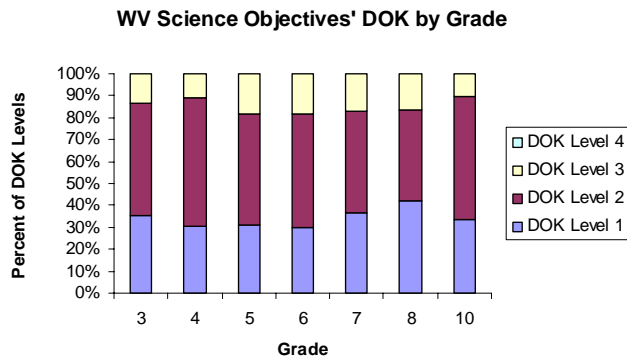
As for the other two content areas, eight reviewers reviewed the objectives under the six science standards and reached consensus on the depth-of-knowledge for each objective. Half of the reviewers were science teachers and curriculum coordinators from West Virginia and half were science educators from other states. Unlike the other two content areas, the depth-of-knowledge levels remained fairly consistent among the four DOK levels over the seven grades analyzed, with about one third of the objectives rated at a DOK level of 1 (Recall and Recognition), one half of the objectives rated at a DOK level 2 (Skills and Concepts), and one-sixth of the objectives rated at a DOK level 3 (Strategic Thinking) (Exhibit 15). What changed over the grades were the scientific topics and ideas that students were expected to know and to do. For example, the grade 3 Objective 3.4.7 expects students to “relate changes in states of matter to changes of temperature.” The grade 8 Objective 8.4.17 expects students to “identify chemical reaction factors that might affect the reaction rates, including catalysts, temperature changes, light energies, and particle size.” The grade 3 objective was rated at DOK Level 2 because it required students to relate changes in states of matter and temperature. The grade 8 objective was rated at DOK Level 1 because it expected students to only identify chemical reaction factors. However, the grade 8 objective required that students know more science, including something about reaction rates, catalysts, temperature changes, and light energies. The relative complexity of the expectations for students at a grade level remained very similar over the grades, but the scientific ideas students were expected to know did increase over time. As a consequence, even though the DOK levels remained flat (Exhibit 16) across the seven grades, the science assessment system showed coherence by requiring students to increase their scientific knowledge and understanding.

The blueprint for the science assessments indicated that the proportion of items among the six standards were to be distributed in the same proportions at each of the

Percent of Objectives by Depth-of-Knowledge Levels for Each Grade, West Virginia Alignment Analysis for Science

Grade	Number of Objs	DOK Level	# of objs by Level	% w/in std by Level
Grade 3	45	1	16	35
		2	23	51
		3	6	13
Grade 4	59	1	18	30
		2	34	57
		3	7	11
Grade 5	48	1	15	31
		2	24	50
		3	9	18
Grade 6	54	1	16	29
		2	28	51
		3	10	18
Grade 7	57	1	21	36
		2	26	45
		3	10	17
Grade 8	60	1	24	40
		2	25	42
		3	10	16
Grade 10	66	1	22	33
		2	37	56
		3	7	10

Exhibit 16
Science DOK Levels for Objectives by Grade



grades with the highest proportion of items (38%) related to Standard 4 (Subject as Concepts) and the next highest (30%) related to Standard 2 (Inquiry). Reviewers found

assessment items that related to each of the six standards for all of the grades. The proportions of items found by the reviewers were similar to the proportions identified in the blueprint, but varied from grade to grade (Exhibit 17), with the highest proportions of items related to Standard 4. In general, the assessment items for each grade were appropriate for the grade. However, reviewers did notice instances, at every grade level, where particular items would obviously have been better aligned with an objective for a preceding grade. However, they did not encounter instances of items being too advanced—i.e., where the item would be better aligned with objectives at a higher grade level than the assessment being analyzed. Thus, the assessment system had some coherence, but reviewers felt it could be improved in specific ways.

There were 50 items on the assessment forms analyzed for six of the seven grades and 52 items on the grade 8 assessment form. At each grade level, the assessment included five open-response items. These items had a value of two points compared to the one-point value given to all of the multiple-choice items. For grades 3-8 all open-response questions had a point value of two points. At grade 10, all of the five open-response items had a point value of four. Thus, the total points possible on the science assessments ranged from 55 (grade 3) to 65 (grade 10). In their computations of the comparison of the assessment and curriculum standards in the alignment analysis, the reviewers took into consideration the point-value given to the item.

Exhibit 17

Percent of Hits (and Point Values) by Standard for Each Grade on the Assessments and on the State’s Blueprint-Designed Range

Science Standards	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
1. History/Nature	5%	8%	3(4)%	6%	6%	4(3)%	3(2)%
Blueprint	8%	8%	8%	8%	8%	8%	8%
2. Inquiry	30(31)%	22(24)%	24(22)%	32(31)%	29(30)%	26(25)%	15(11)%
Blueprint	30%	30%	30%	30%	30%	30%	30%
3. Unifying Themes	7%	4%	7%	4%	5%	3%	9(7)%
Blueprint	8%	8%	8%	8%	8%	8%	8%
4. Subject Matter/Concepts	48(44)%	51(50)%	50(52)%	48(49)%	44%	53(59)%	61(72)%
Blueprint	38%	38%	38%	38%	38%	38%	38%
5. Design/Applications	4%	4(3)%	7%	4%	3%	5(3)%	3(2)%
Blueprint	8%	8%	8%	8%	8%	8%	8%
6. Personal/Social	6(9)%	11%	9(8)%	6%	13(12)%	9(7)%	9(6)%
Blueprint	8%	8%	8%	8%	8%	8%	8%

Overall, the alignment of the science curriculum standards for the seven grades and the corresponding assessments are in need of some improvement in order to achieve assessment of the full spectrum of the standards. The alignment between the standards and the assessments was seen as adequate on all four criteria for the standard given the highest importance, Standard 4 (Subjective Matter and Concepts).

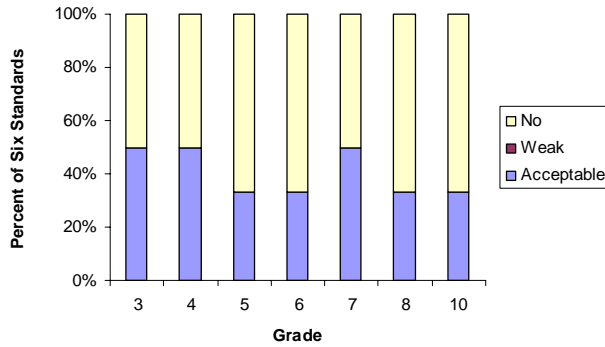
The assessments at each grade level primarily measured content related to two of the six curriculum standards—Standard 2 (Unifying Themes) and Standard 4 (Subject Matter and Concepts). Reviewers found that all of the other standards had at least one or two items that measured content related to those standards, but the number of the related items fell short of the required six items used as the acceptable basis for reporting about a student’s performance on a standard. Across the grade levels, Standard 5 (Design and Applications) followed by Standard 1 (History and Nature) had the fewest number of corresponding items—on the average, two or three items. It was clear to the reviewers and from the two analyses, one in July and one in November, that in the November analysis there was some increase in the number of items that related to these less emphasized standards and that there was improvement in the revised assessment reviewed in November. Consistent with the low number of items related to these standards, over half of the items on the assessments that corresponded to Standards 2, 5, and 6 had a DOK level that was below that of the corresponding standard. In addition, it was found that across the grades an increasing number of the standards and the assessments failed to meet an acceptable level on the Range-of-Knowledge Correspondence criterion; that is, less than half of the objectives under a standard had at least one item measuring related content.

By design, a high proportion (70% to 80%) of the items corresponded to objectives under the two standards (Standards 2 and 4) of the six considered the most important. As a consequence, reviewers coded an insufficient number of items, less than six, as corresponding to the other four standards. For three grades, only three standards met the acceptable level of having six or more items (Exhibit 18). For the other three grades, only two standards met this level. The difference in the number of corresponding items among the six science standards is reasonable in light of the expectations of the standards. Many of the objectives under Standard 5 (Design and Applications) expect students to do some activity that is difficult to measure on an on-demand assessment. For example, a grade 6 objective (6.5.1) under Standard 5 states, “Given a set of attributes, produce a product or process and cite how design priorities (e.g., space, safety) and available materials impact the outcome.” At grade 10, Objective 10.5.2 directs students to “research and design solutions to a personal or a societal problem created by technology.” The assessments across the grades did include from two to four items related to Standard 5. At least some content related to design and applications was assessed, even though much of what the standard expected was that students perform a task. Two other standards, Standard 1 (History/Nature) and Standard 3 (Unifying Themes), both related to students’ understanding the nature of science and its role over time. If these standards were combined into one standard, as is the case in some sets of standards, then the assessments would have met the acceptable level of six items. A very reasonable case can be made for the West Virginia science standards not being required to have six items for each of the standards. The assessment at each grade did include at least some items measuring content related to each of the six standards. Also, the two standards deemed most important had the greatest number of items. Thus, the assessment results will produce adequate data for making some judgments on students’ knowledge of scientific concepts and inquiry while also producing at least some data on students’ knowledge of the nature of science and its use in society. Even though the items were not evenly

distributed among the six standards, the number of assessment items across the standards was sufficient to report on students' attainment of the important science constructs identified by the state.

Exhibit 18

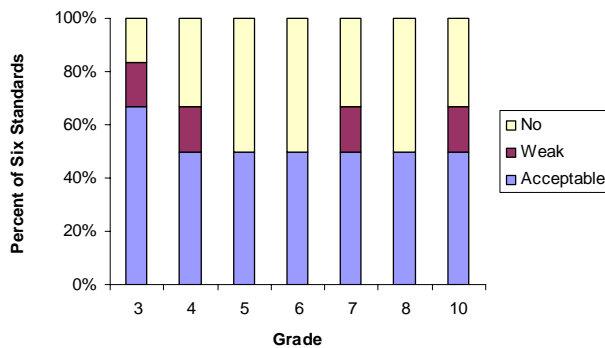
Percent of Six Science Standards with An Acceptable Number of Items by Grade



Half of the six science standards and the assessments for each of the grades had an acceptable level on the content complexity, or Depth-of-Knowledge Consistency criterion, of 50% or more of the items with a DOK level at or above the corresponding objective. The three standards that only weakly met, or failed to meet, this acceptable level were those standards that expected students to carry out some form of inquiry, production, or investigation (standards 2, 5, and 6). These are the only standards for which reviewers rated objectives at a DOK level 3 (Strategic Thinking). Reviewers rated nearly all of the assessment items for each of the grades as having a DOK level of 1 or 2. The low percentage of assessment items with a DOK level of 3 is a shortcoming of the assessments and one area in which some improvement is needed.

Exhibit 19

Percent of Six Science Standards by An Acceptable Level of Depth-of-Knowledge Consistency by Grade

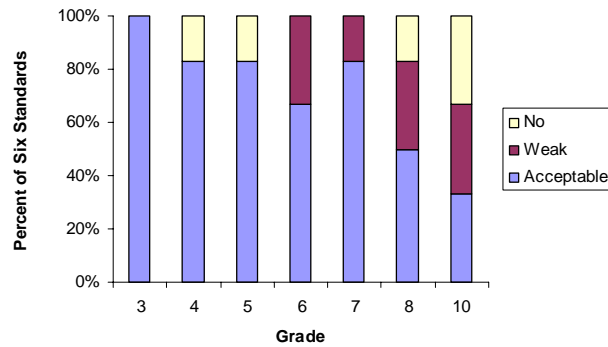


The range of content covered by the assessments was generally good except for grades 8 and 10 (Exhibit 20). Over half of the objectives had a least one corresponding

item for five of the six standards for four grades. For two of the grades (grades 6 and 8), this acceptable level was at least weakly met for five of the six standards. Only at grade 10 were the assessment items less than adequately distributed among the objectives for over half of the standards. The Range-of-Knowledge Correspondence criterion was met for Standard 4 (Subject Matter and Concepts) for each of the grades, even though the number of objectives under this standard ranged from 20 to 40 objectives—in part, because of the large number of items for each grade that corresponded to Standard 4. Thus, with the exception of grade 10, the assessments covered an adequate range of content.

Exhibit 20

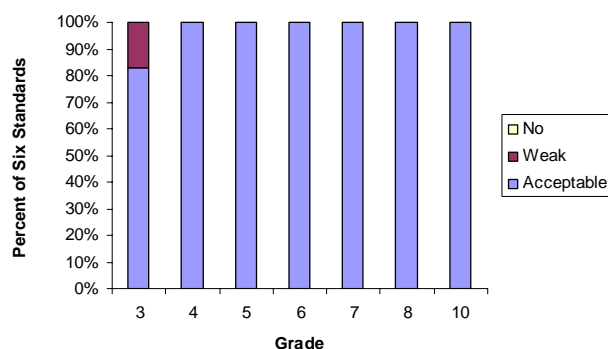
Percent of Six Science Standards with An Acceptable Range-of-Knowledge Correspondence by Grade



The items on the science assessments were distributed evenly enough among the objectives under the standards to achieve a high degree of balance. An acceptable level on the Balance-of-Representation criterion was fully met for six of the seven grades and nearly fully met for the other grade (Exhibit 21).

Exhibit 21

Percent of Six Science Standards by Acceptable Level of Balance of Representation by Grade



Overall, the alignment of the science curriculum standards for the seven grades and the corresponding assessments had areas in need of improvement, but provided an assessment system that fulfilled its purposes. The assessments at each grade level primarily measured content related to two of the six curriculum standards—Standard 2 (Inquiry) and Standard 4 (Subject Matter and Concepts). Reviewers found that all of the other standards at least had one or two items that measured content related to those standards, but the number of the related items fell short of the requirement of six items (used as the acceptable level) to be able to report on a student’s performance on a standard. Considering the design and the intent of the assessment, this was not deemed as a critical failing in alignment. Across the grade levels, Standard 5 (Design and Applications), followed by Standard 1 (History and Nature), had the fewest number of corresponding items—on the average, two or three items. Corresponding to the low number of items related to these standards, over half of the items on the assessments that corresponded to Standards 2, 5, and 6 had a DOK level that was below that of the corresponding objective. The need for some additional items at a higher DOK level compared to the level of the corresponding objectives was considered the most critical alignment issue. Even though across the grades an increasing number of the standards and the assessments failed to meet an acceptable level on the Range-of- Knowledge Correspondence criterion, this was not so significant, except at grade 10. It was very apparent that Standard 4 (Subject Matter and Concepts) is regarded as a dominant standard, as indicated by the number of objectives specified under the standard in relation to the other standards. Over the seven grades, the number of objectives under Standard 4 ranges from 23 (grade 3) to 40 (grade 10); the number of objectives under any of the other five standards ranges from 2 to 9. Even with the large number of objectives, there was very good alignment between Standard 4 and the assessment at each of the grades. Overall, the alignment of the science standards and assessment was judged to be strong for Standard 4, moderate for Standard 2, and weak for the other four standards, mainly because of the distribution of the assessment items among the six standards and the need for some DOK level 3 items for three of the standards.

The Alignment Process

Three groups of eight content experts, four from West Virginia and four from other states, analyzed the agreement between the West Virginia academic standards and the WESTEST. One group analyzed the agreement in each of three content areas—mathematics, reading language arts, and science. The analyses were done in two stages, once in July, 2003, and once in November, 2003. During the intervening period, state staff worked with the testing vendor to improve the standards and assessment alignment based on the information provided in the first analysis. In November, three detailed reports were produced, one for each content area (Webb, 2003a and b; 2004). At each institute, reviewers were trained in the process of identifying the four depth-of-knowledge levels (Appendix A). Reviewers reached consensus on the depth-of-knowledge levels for each objective under each standard for a content area. Then reviewers coded the depth-of-knowledge level for each assessment item, one primary objective, and up to two secondary objectives. This process produced results on four criteria—Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Correspondence, and Balance-of-Representation (Appendix B). Codings across the eight reviewers were averaged to produce a value for each criterion. These values were compared to a predefined acceptable level based on a specific rationale. Intraclass correlations were computed to determine the agreement among reviewers in assigning a depth-of-knowledge level to the assessment items (Appendix C). Reviewers also made note of any items with a source-of-challenge that could cause results from an item to be misinterpreted and other comments specific to an item.

Conclusions

These analyses of the West Virginia academic standards and WESTEST assessments found there was strong alignment in mathematics, acceptable alignment in reading language arts, and alignment on the most significant content in science as identified by the state. A progression in complexity in the expectations for students appropriate to the content area was observed in each content area. For a content area, each grade had the same general standards, but the expectations progressed across the grades. In mathematics and reading, this was evident in an increase in the percentage of objectives with higher depth-of-knowledge levels in the higher grades. In science, this progression was evident in the increasing sophistication of scientific content students were expected to know. As such, the assessment system had coherence in that students' progress could be tracked from one grade to the next and students were required to have greater sophistication in the content area from grade to grade.

In mathematics, the analyses indicated that there were a sufficient number of items to judge attainment of each of the five standards, and these items were adequately distributed among the objectives to measure a range in the required knowledge under each standard. For nearly all of the standards across the seven grades, the items matching objectives were distributed fairly evenly. The depth-of-knowledge (DOK) levels of a small number of items, mainly those relating to one standard, were lower than the level of corresponding objectives for five of the seven grades. However, this could readily be remedied and was judged not to be critical. Overall, the mathematics assessment measured the breadth and depth of the academic standards along with an adequate range

within each standard. The emphasis on the assessments was evenly distributed among the objectives under each standard with no one objective being over-emphasized. Thus, the results of the analysis indicated that the West Virginia academic standards and the WESTEST in mathematics were in alignment.

In reading language arts, the alignment between the West Virginia academic standards and the WESTEST was, in general, reasonable, given the constraints on the assessment. One standard, Listening/Speaking/Viewing, could not be assessed on the on-demand assessment. There were an adequate number of items on the assessments at all seven grade levels to measure students' attainment of the other two standards, Reading and Writing. For all of the grades, except for the grade 10 Reading standard, the items were at a depth-of-knowledge level that adequately matched the depth-of-knowledge levels of the corresponding objectives. At four of the grades, an adequate proportion of the objectives had at least one corresponding item for the Range-of-Knowledge Correspondence criterion to be met for both the Reading and the Writing standards. The distribution of items could be improved for the remaining three grades, but because other objectives were assessed on other forms, the lack of attainment of the Range-of-Knowledge criterion for these three grades was not considered critical. By design, some areas under the Reading and Writing standards were emphasized more than other areas on the assessment. Because an acceptable alignment was generally achieved on the other criteria across the grades, this imbalance was not viewed as significant. As such, the assessments adequately measured the depth and breadth of the academic standards, at an appropriate depth of knowledge, and with the emphasis desired by the state.

In science, analysis of the alignment of the academic standards for the seven grades and the corresponding assessments showed areas in need of improvement, but found that the assessment system fulfilled its purposes. The assessments at each grade level primarily measured content related to two of the six curriculum standards—Standard 2 (Inquiry) and Standard 4 (Subject Matter and Concepts). These standards, which cover the most significant content, were indicated on the state's assessment blueprint as being the most important. Reviewers found that each of the other standards at least had one or two items that measured content related to those standards, but that the number of the related items fell short of the required six (used as the acceptable level) to be able to report on a student's performance on a standard. Considering the design and the intent of the assessment, this was not deemed a critical failing in alignment. Corresponding to the low number of items related to the three standards, over half of the items on the assessments that corresponded to these standards had a low depth-of-knowledge level in relation to the corresponding objective. The need for some additional items at a higher DOK level compared to the level of the corresponding objectives was considered the most critical alignment issue.

References

- Subkoviak, M. J. (1988). A practioner's guide to computation and interpretation of reliability indices for mastery texts. *Journal of Educational Measurement*, 25(1), 47-55.
- Webb, N. L. (2003a). *Alignment analysis of mathematics standards and assessments using the operational forms, West Virginia, Grades 3, 4, 5, 6, 7, 8, and 10*. Madison, WI.
- Webb, N. L. (2003b). *Alignment analysis of reading language arts standards and assessments using the operational forms, West Virginia, Grades 3, 4, 5, 6, 7, 8, and 10*. Madison, WI.
- Webb, N. L. (2004). *Alignment analysis of science standards and assessments using the operational forms, West Virginia, Grades 3, 4, 5, 6, 7, 8, and 10*. Madison, WI.

Appendices

A. Depth-of-Knowledge Definitions

B. Four Alignment Criteria Used in the Analysis

C. Reviewer Agreement in Coding Depth-of-Knowledge Levels

Appendix A

Depth-of-Knowledge Definitions

Mathematics

Level 1 (Recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. In science, a simple experimental procedure, including one or two steps, should be coded as Level 1. Other key words that signify a Level 1 include “identify,” “recall,” “recognize,” “use,” and “measure.” Verbs such as “describe” and “explain” could be classified at different levels, depending on what is to be described and explained.

Level 2 (Skills/Concepts) includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as “explain,” “describe,” or “interpret” could be classified at different levels, depending on the object of the action. For example, interpreting information from a simple graph, which requires reading information from the graph, also is a Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is a Level 3. Caution is warranted in interpreting Level 2 as only skills because some reviewers will interpret skills very narrowly, as primarily numerical skills; such interpretation excludes from this level other skills such as visualization skills and probability skills, which may be more complex simply because they are less common. Other Level 2 activities include explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Other Level 3

activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

Level 4 (Extended Thinking) requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student were to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. At this highest level, students should be required to make several connections—relate ideas *within* the content area, or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Reading Language Arts

Reading

Reading Level 1. Level 1 requires students to receive or recite facts, or to use simple skills or abilities. Oral reading that does not include analysis of the text, as well as basic comprehension of a text, is included. Items require only a shallow understanding of the text presented and often consist of verbatim recall from text, or simple understanding of a single word or phrase. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Support ideas by reference to details in the text.
- Use a dictionary to find the meanings of words.
- Identify figurative language in a reading passage.

Reading Level 2. Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Intersentence analysis of inference is required. Some important concepts are covered, but not in a complex way. Standards and items at this level may include words such as “summarize,” “interpret,” “infer,” “classify,” “organize,” “collect,” “display,” “compare,” and “determine whether fact or opinion.” Literal main ideas are stressed. A Level 2 assessment item may require students to apply skills and concepts that are covered in Level 1. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Use context cues to identify the meaning of unfamiliar words.

- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

Reading Level 3. Deep knowledge becomes a greater focus at Level 3. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge. Items may also involve more superficial connections between texts. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Determine the author's purpose and describe how it affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

Reading Level 4. Higher-order thinking is central and knowledge is deep at Level 4. The standard or assessment item at this level will probably be an extended activity, with additional time provided for completing it. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking. Students take information from at least one passage of a text and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts. Some examples that represent, but do not constitute all of, Level 4 performance are:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

Writing

Writing Level 1. Level 1 requires the student to write or recite simple facts. The focus of this writing or recitation is not on complex synthesis or analysis but on basic ideas. The students are asked to list ideas or words, as in a brainstorming activity prior to written composition; are engaged in a simple spelling or vocabulary assessment; or are asked to write simple sentences. Students are expected to write and speak using the conventions of Standard English. This includes using appropriate grammar, punctuation, capitalization, and spelling. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Use punctuation marks correctly.
- Identify Standard English grammatical structures and refer to resources for correction.

Writing Level 2. Level 2 requires some mental processing. At this level, students are engaged in first-draft writing, or brief extemporaneous speaking for a limited number of purposes and audiences. Students are expected to begin connecting ideas, using a simple organizational structure. For example, students may be engaged in note-taking, outlining, or simple summaries. Text may be limited to one paragraph. Students demonstrate a basic understanding and appropriate use of such reference materials as a dictionary, thesaurus, or Web site. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Construct compound sentences.
- Use simple organizational strategies to structure written work.
- Write summaries that contain the main idea of the reading selection and pertinent details.

Writing Level 3. Level 3 requires some higher-level mental processing. Students are engaged in developing compositions that include multiple paragraphs. These compositions may include complex sentence structure and may demonstrate some synthesis and analysis. Students show awareness of their audience and purpose through focus, organization, and the use of appropriate compositional elements. The use of appropriate compositional elements may include addressing chronological order in a narrative, or including supporting facts and details in an informational report. At this stage, students are engaged in editing and revising to improve the quality of the composition. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Support ideas with details and examples.
- Use voice appropriate to purpose and audience.
- Edit writing to produce a logical progression of ideas.

Writing Level 4. Higher-level thinking is central to Level 4. The standard at this level is a multiparagraph composition that demonstrates the ability to synthesize and analyze complex ideas or themes. There is evidence of a deep awareness of purpose and audience. For example, informational papers include hypotheses and supporting evidence. Students are expected to create compositions that demonstrate a distinct voice and that stimulate the reader or listener to consider new perspectives on the ideas and themes addressed. An example that represents, but does not constitute all of, Level 4 performance is:

- Write an analysis of two selections, identifying the common theme and generating a purpose that is appropriate for both.

Science

Level 1 (Recall and Reproduction) is the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple science process or

procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. A “simple” procedure is well-defined and typically involves only one step. Verbs such as “identify,” “recall,” “recognize,” “use,” “calculate,” and “measure” generally represent cognitive work at the recall and reproduction level. Simple word problems that can be directly translated into and solved by a formula are considered Level 1. Verbs such as “describe” and “explain” could be classified at different DOK levels, depending on the complexity of what is to be described and explained.

A student answering a Level 1 item either knows the answer or does not: that is, the answer does not need to be “figured out” or “solved.” In other words, if the knowledge necessary to answer an item automatically provides the answer to the item, then the item is at Level 1. If the knowledge necessary to answer the item does not automatically provide the answer, the item is at least at Level 2. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Recall or recognize a fact, term, or property.
- Represent in words or diagrams a scientific concept or relationship.
- Provide or recognize a standard scientific representation for simple phenomenon.
- Perform a routine procedure, such as measuring length.

Level 2 (Skills and Concepts) includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is *more complex* than in Level 1. Items require students to make some decisions as to how to approach the question or problem. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply *more than one step*. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Some action verbs, such as “explain,” “describe,” or “interpret,” could be classified at different DOK levels, depending on the complexity of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is a Level 2. An item that requires interpretation from a complex graph, such as making decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated, is at Level 3. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Specify and explain the relationship between facts, terms, properties, or variables.
- Describe and explain examples and non-examples of science concepts.
- Select a procedure according to specified criteria and perform it.
- Formulate a routine problem, given data and conditions.
- Organize, represent, and interpret data.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity results not only from the fact that there could be multiple answers, a possibility for both Levels 1 and 2, but because the multi-step task requires more demanding reasoning. In most instances, requiring students to explain their thinking is at Level 3; requiring a very simple explanation, or a word or two, should be at Level 2. An activity that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Experimental designs in Level 3 typically involve more than one dependent variable. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems. Some examples that represent, but do not constitute all of, Level 3 performance, are:

- Identify research questions and design investigations for a scientific problem.
- Solve non-routine problems.
- Develop a scientific model for a complex situation.
- Form conclusions from experimental data.

Level 4 (Extended Thinking) requires high cognitive demands and complexity. Students are required to make several connections—relate ideas within the content area or among content areas—and to select or devise one approach among many alternatives on how the situation can be solved. Many on-demand assessment instruments will not include any assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. “Develop generalizations of the results obtained and the strategies used and apply them to new problem situations,” is an example of a grade 8 objective that is a Level 4. Many, but not all, performance assessments and open-ended assessment activities requiring significant thought will be Level 4.

Level 4 requires complex reasoning, experimental design and planning, and probably will require an extended period of time either for the science investigation required by an objective, or for carrying out the multiple steps of an assessment item. However, the extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be a Level 4. Some examples that represent, but do not constitute all of, a Level 4 performance are:

- Based on provided data from a complex experiment that is novel to the student, deduct the fundamental relationship between several controlled variables.
- Conduct an investigation, from specifying a problem to designing and carrying out an experiment, to analyzing its data and forming conclusions.

Appendix B

Four Alignment Criteria Used in the Analysis

Categorical Concurrence

An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. *The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content from each standard. The analysis assumed that the assessment had to have at least six items measuring content from a standard in order for an acceptable level of categorical concurrence to exist between the standard and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors have to be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. This indicates that about 63% of the group would be consistently classified as masters or nonmasters if two equivalent test administrations were employed. The agreement coefficient would increase if the cutoff score is increased to one standard deviation from the mean to .77 and, with a cutoff score of 1.5 standard deviations from the mean, to .88. Usually states do not report student results by standards, or require students to achieve a specified cutoff score on subscales related to a standard. If a state did do this, then the state would seek a higher agreement coefficient than .63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a standard and as a basis for making some decisions about students' knowledge of that standard. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the standard, as judged in this analysis, at least 50% of the items corresponding to an objective has to be at or above the level of knowledge of the objective: 50%, a conservative cutoff point, is based on the assumption that a minimal passing score for any one standard of 50% or higher would require the student to

successfully answer at least some items at or above the depth-of-knowledge level of the corresponding objectives. For example, assume an assessment included six items related to one standard and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding objectives, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one objective. Some leeway was used in this analysis on this criterion. If a standard had between 40% to 50% of items at or above the depth-of-knowledge levels of the objectives, then it was reported that the criterion was “weakly” met.

Range-of-Knowledge Correspondence

For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of objectives within the standard with one related assessment item/activity. Fifty percent of the objectives for a standard had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students’ knowledge should be tested on content from over half of the domain of knowledge for a standard. This assumes that each objective for a standard should be given equal weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one objective, the requirement that assessment items be related to more than 50% of the objectives for a standard increases the likelihood that students will have to demonstrate knowledge on more than one objective per standard to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the objectives. However, any restriction on the number of items included on the test will place an upper limit on the number of objectives that can be assessed. Range-of-knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of standards and a large number of objectives. If 50% or more of the objectives for a standard had a corresponding assessment item, then the range-of-knowledge criterion was met. If between 40% and 50% of the objectives for a standard had a corresponding assessment item, the criterion was “weakly” met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion only considers the number of objectives within a standard hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these objectives. *The balance-of-*

representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another. An index is used to judge the distribution of assessment items. This index only considers the objectives for a standard that have at least one hit—i.e., one related assessment item per objective. The index is computed by considering the difference in the proportion of objectives and the proportion of hits assigned to the objective. An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the objectives for the given standard. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the objectives hit. Depending on the number of objectives and the number of hits, a unimodal distribution (most items related to one objective and only one item related to each of the remaining objectives) has an index value of less than .5. A bimodal distribution has an index value of around .55 or .6. Index values of .7 or higher indicate that items/activities are distributed among all of the objectives at least to some degree (e.g., every objective has at least two items) and is used as the acceptable level on this criterion. Index values between .6 and .7 indicate the balance-of-representation criterion has only been “weakly” met.

Appendix C
Reviewer Agreement in Coding Depth-of-Knowledge Levels
Mathematics

Intraclass Correlation Among Eight Reviewers in Assigning Item Depth-of-Knowledge Level for Mathematics November 2003

Grade	Intraclass Correlation
3	0.930
4	0.928
5	0.935
6	0.916
7	0.875
8	0.942
10	0.900

Reading Language Arts

Intraclass Correlation Among Eight Reviewers in Assigning Item Depth-of-Knowledge Level for Reading Language Arts November 2003

Grade	Intraclass Correlation
3	.840
4	.847
5	.857
6	.764
7	.756
8	.788
10	.764

Science

Intraclass Correlation Among Eight Reviewers in Assigning Item Depth-of-Knowledge Level for Science November 2003

Grade	Intraclass Correlation
3	0.835
4	0.734
5	0.843
6	0.866
7	0.826
8	0.820
10	0.865

