

REPORT

Alignment Analysis of Science Standards and Pre-Field Test Assessments

**West Virginia
Grades 3-9
2008**

Norman L. Webb

October 12, 2008

REPORT

Alignment Analysis of Science Standards and Pre-Field Test Assessments

**West Virginia
Grades 3-9
2008**

Norman L. Webb

October 12, 2008

Acknowledgements

Reviewers:

Grades 3-6

Sara Christopherson	Group Leader	WI
Jim Leidel		WI
Gwen Pollock		IL
Sherry Baker		WV
Kim Shomo		WV
Kimberly Fulton		WV

Grades 6-9

Jim Woodland	Group Leader	NE
Douglas Johnson		WI
Katya Denisova		MD
Charles Vest		WV
Deena Young		WV
Jan Click		WV

The State of West Virginia funded this analysis. Brenda West, Assistant Director, Office Assessment and Accountability, was the main contact person for the West Virginia Department of Education. Jan Barth, Special Assignment, Superintendent's Office and Director of Office of Assessment and Accountability, had overall responsibility for the study.

Table of Contents

Executive Summary	v
Introduction.....	1
Alignment Criteria Used for This Analysis	3
Categorical Concurrence.....	3
Depth-of-Knowledge Consistency.....	4
Range-of-Knowledge Correspondence.....	6
Balance of Representation	6
Source of Challenge.....	7
Findings.....	7
Standards.....	7
Alignment of Curriculum Standards and Assessments.....	9
Source-of-Challenge Issues and Reviewers’ Comments	21
Reliability Among Reviewers.....	21
Summary	22
References.....	23
Appendix A	
West Virginia Grades 3-9 Science Standards and Group Consensus DOK Values	
Appendix B	
Data Analysis Tables West Virginia Grades 3-9 WESTEST2 Forms 1 and 2 Science 2008	
Appendix C	
Reviewers’ Notes and Source-of-Challenge Comments West Virginia Grades 3-9 WESTEST2 Forms 1 and 2 Science 2008	
Appendix D	
Debriefing Summary Notes West Virginia Grades 3-9 WESTEST2 Forms 1 and 2 Science 2008	
Appendix E	
Data Analysis Tables West Virginia Grade 6 TerraNova Form Science 2008	

Appendix F
Reviewers' Notes and Source-of-Challenge Comments West Virginia Grade 6 TerraNova
Forms Science 2008

Appendix G
Debriefing Summary Notes West Virginia Grade 6 TerraNova Forms Science 2008

Executive Summary

A three day alignment institute was held September 17-19, 2008 in Charleston, West Virginia to analyze the pre-field test forms of the WESTEST2 and TerraNova with the West Virginia 21st century science standards. Two groups of six reviewers each participated in the institute. One group analyzed assessments and standards for grades 3-6 and one group analyzed these documents for grades 6-8 and grade 9 physical science. Both groups independently analyzed the alignment between the standards and grade 6 Form 1. Half of the reviewers were from West Virginia and half were from other states. The reviewers included science education content experts, state science supervisors, and science teachers. Two forms of the WESTEST2 assessment for each grade and one TerraNova grade six form were analyzed.

For nearly all of the fifteen science forms analyzed, the alignment between the WESTEST2 assessments and the West Virginia 21st century science standards was acceptable. The science standards and two WESTEST2 forms were found to be fully aligned (grade 3 Form 2 and grade 4 Form 2). For the other WESTEST2 science forms some minor alignment issue was detected. All of the WESTEST2 forms for grades 5 through 9 had fewer than six items that targeted objectives under Standard III (Application of Science). The lower number of items for Standard III was generally accompanied by too many items with too low of a DOK level or a lack of range. The TerraNova grade 6 assessment and the grade 6 standards needed major improvement.

Reviewers commented that the science items were generally reasonable. The one repeated comment by reviewers in the grade 3-6 group was that some of the items were more of reading items where the students were required to infer from the prompt the science rather than recall or apply conceptual knowledge. More specific comments on individual items are included in the appendices. Overall, the alignment for science was at least acceptable with fewer than five items needed to be replaced or added to attain full alignment as summarized in the table on the next page.

Summary Table

Percent of West Virginia Mathematics Standards with Acceptable Level on Each Alignment Criteria for Grade 3-8 and 9 Physical Science for WESTEST2 Analysis

Grade	<i>Categorical Concurrence</i> (six or more items)	<i>Depth-of-Knowledge Consistency</i> (50% at/above)	<i>Range of Knowledge</i> (50% of objectives)	<i>Balance of Representation</i> (without possible weakness)	<i>Estimated Range of Items per to be Added or Replaced for Full Alignment</i>
3 Form 1	100	100	33	100	2
3 Form 2	100	100	100	100	0
4 Form 1	100	100	67	100	2
4 Form 2	100	100	100	100	0
5 Form 1	67	100	100	100	1
5 Form 2	67	100	100	100	1
6 Form 1	67	100	67	100	4
6 Form 2	67	67	67	67	5
7 Form 1	67	100	100	100	3
7 Form 2	67	100	100	100	3
8 Form 1	67	67	67	100	3
8 Form 2	67	100	100	100	1
9 Form 1	67	67	67	100	2
9 Form 2	67	33	67	100	2
6 TerraNova	67	67	0	100	13

Categorical Concurrence >6 items
 Depth-of-Knowledge >50% with DOK level the same or higher than level of corresponding Objectives
 Range-of-Knowledge >70% of objectives under a standard
 Balance of Representation A possible weakness if one or more objectives with a relative large number of items (e.g. five or more than the objective with the next highest number of items)

Alignment Analysis of Science Standards and Pre-Field Test Assessments

West Virginia Grades 3-9 2008

Norman L. Webb

Introduction

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute of any one of these two system components. Alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997).

A three-day Alignment Analysis Institute was conducted September 17-19, 2008, in Charleston, West Virginia. Two groups of six reviewers each analyzed the relationship between two WESTEST2 assessment forms and one TerraNova form for each grade from grade 3 through grade 9 (physical science). The assessments are being developed for administration in the 2008-09 school year. Both groups analyzed one form of the WESTEST 2 assessment for grade 6. Then the grades 3-6 group analyzed the grade 6 TerraNova form and all of the forms for grades 3-5. The grades 6-9 group analyzed the WESTEST2 Form 2 for grades 6 and all forms for grades 7-9. Each group included three reviewers from West Virginia and three reviewers from other states. The science reviewers included science content experts, state science supervisors, and science teachers.

The 21st Century West Virginia Content Standards and Objectives (CSOs) used the terminology of *content standards* and *objectives* in its science content expectations. Standards were the broad content requirements across all grades. There were three standards for each grade and the grade 9 physical science—Nature of Science, Content of Science, and Applications of Science. Objectives specified in greater detail under a standard what students are to know and do. Each standard had from six to 36 objectives. Data for this analysis were entered for the objectives and reported out at the standards level.

As part of the alignment institute, reviewers were trained to identify the depth-of-knowledge of the objectives and assessment items. This training included reviewing the definitions of the four depth-of-knowledge (DOK) levels and reviewing examples of each. Then the reviewers participated in 1) a review of the depth-of-knowledge levels pre-assigned to the objectives and 2) individual analyses of the assessment items. In reviewing the DOK levels of the objectives, reviewers were instructed not to change any of the DOK values. If the reviewers disagreed with the DOK level assigned to an objective, they were to make a note of this disagreement, but not to change any of the DOK values. The science reviewers did not find any DOK levels they disagreed with and did not change any values. Following individual analyses of the items, reviewers participated in a debriefing discussion if time allowed in which they assessed the degree to which they had coded particular items or types of content to the objectives. Because of the large volume of work, for some assessment parts, three or four reviewers conducted the analysis on a part.

Two test forms for the WESTEST2 assessment to be used for the first time in spring 2008-2009 were analyzed along with one TerraNova form for grade 6. Items were written by WV teachers and CTB staff and were included in the item bank alignment analysis conducted in April 2008.

To derive the results from the analysis, the reviewers' responses were averaged. Any variance among reviewers was considered legitimate, with the true depth-of-knowledge level for the item falling somewhere between two or more assigned values. Such variation could signify a lack of clarity in how the standards and objectives were written, the robustness of an item that can legitimately correspond to more than one objective, and/or a depth of knowledge that falls in between two of the four defined levels. Reviewers were allowed to identify one assessment item as corresponding to up to three objectives—one primary hit (objective) and up to two secondary hits. However, reviewers could only code one depth-of-knowledge level to each assessment item, even if the item corresponded to more than one objective.

Reviewers were instructed to focus primarily on the alignment between the state standards and assessments. However, reviewers were encouraged to offer their opinions on the quality of the standards, or of the assessment activities/items, by writing a note about the item (see Appendices C and F). Reviewers could also indicate whether there was a source-of-challenge issue with an item—i.e., a problem with the item that might cause the student who knows the material to give a wrong answer, or enable someone who does not have the knowledge being tested to answer the item correctly.

The results produced from the institute pertain only to the issue of alignment between the West Virginia state standards and the items from the item bank. Note that this alignment analysis does not serve as external verification of the general quality of the state's standards or assessment items. Although reviewers did make a number of suggestions on issues with items or how items could be improved. Only the degree of alignment is discussed in the results. For these results, the means of the reviewers' coding were used to determine whether the alignment criteria were met. When reviewers did

vary in their judgments, the means lessened the error that might result from any one reviewer's finding. Standard deviations are reported in the tables provided in the appendices, which give one indication of the variance among reviewers.

The study addressed specific criteria related to the content agreement between the state standards and grade-level assessment items. Four criteria received major attention: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation.

Alignment Criteria Used for This Analysis

This analysis judged the alignment between the standards and the assessments on the basis of four criteria. Information is also reported on the quality of items by identifying items with sources-of-challenge and other issues. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had met the standards.

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 one assessment form had to have at least six items for 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 an assessment form and the standard, as judged in this analysis, at least 50% of the items corresponding to a standard had to be at or above the level of knowledge of the standard: 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 standard. 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 standards, 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 standard. Some leeway was used in this analysis on this criterion. If a standard had between 40% and 50% of items at or above the depth-of-knowledge levels of the standards, then it was reported that the criterion was “weakly” met.

Interpreting and assigning depth-of-knowledge levels to both standards within strands and assessment items are essential requirements of alignment analysis. These descriptions help to clarify what the different levels represent in 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 (e.g. 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.

Level 2 (Skills and Concepts) 2 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 at 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.

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 does *not* result 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.

Level 4 (Extended Thinking). Tasks at Level 4 have high cognitive demands and are very complex. Students are required to make several connections—relate ideas within the content area or among content areas—and have 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 at 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 at Level 4.

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 form 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 benchmark 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 need to 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 correspondence 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.

Source-of-Challenge Criterion

The source-of-challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted science objective, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a source-of-challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed.

Findings

Standards

The DOK levels of the standards assigned prior to this alignment analysis were used. Reviewers reviewed the DOK levels of the objectives and could comment on the level assigned, but did not change any. The DOK levels by grade are summarized in Table 1. The DOK level for each objective is reported in Appendix A. The majority of the DOK levels of the WV Content Standards and Objectives (CSOs) were assigned a DOK level 1 (recall and recognition) and level 2 (skills and concepts). Although there was some decrease in the proportion of objectives assigned a DOK level 1, the distribution of the objectives by DOK levels remained similar across the grades with an increase in proportion of objectives assigned a DOK level 3 or 4 at grades 8 and 9.

Table 1

Percent of Grade-level Expectations by Depth-of-Knowledge (DOK) Levels for Grades 3–8 Science and 9 Physical Science, West Virginia Alignment Analysis for Science 2008 Pre-Field Test Study

Grade	Total Number of Objectives	DOK Level	Number of objectives by Level	Percent within Standard by Level
3	45	1	26	57
		2	16	35
		3	3	6
4	59	1	29	49
		2	23	38
		3	7	11
5	44	1	16	36
		2	20	45
		3	8	18
6	47	1	19	40
		2	22	46
		3	6	12
7	52	1	28	53
		2	18	34
		3	6	11
8	47	1	19	40
		2	18	38
		3	9	19
		4	1	2
9 Physical Science	39	1	13	33
		2	17	44
		3	8	20
		4	1	3

If no particular objective is targeted by a given assessment item, reviewers are instructed to code the item at the level of a learning goal or a standard. This coding to a generic objective sometimes indicates that the item is inappropriate for the grade level. However, if the item is grade-appropriate, then this situation may instead indicate that there is a part of the content not expressly or precisely described in the objectives. These items may highlight areas in the objectives that should be changed, or made more precise. Table 2 displays the assessment items coded to generic objectives by more than one reviewer. Two or more reviewers only coded a total of seven items to generic objectives—three grade 4 items, one grade five item, one grade 6 items, and two grade 7 items. For example, reviewers judged that Item 25 on Grade 5 Form 1 required students to know about how the role of minerals in the body, knowledge that was not found in the standard. The students could answer the question simply by reading the prompt making the item a reading comprehension item rather than science item. The low number of

WESTEST2 science items coded to generic objectives indicates that the science items explicitly targeted the West Virginia objectives. The high number of TerraNova items for grade 6 assigned to generic standards indicates some misfit. Reviewers' reasons for assigning items to generic objectives can be found in Appendices C and F.

Reviewers' debriefing comments also highlight some ambiguities in the objectives. These comments can be found in Appendix D.

Table 2
*Items Coded to Generic Objectives by More Than One Reviewer, West Virginia
 Alignment Analysis for Science, Grades 3-9 2008*

Grade	Generic Objective	Assessment Item (Number of Reviewers)
4 Form 1	SC.S.4.2	13(2)
4 Form 2	SC.S.4.2	13(2); 27(2)
5 Form 1	SC.S.5.2	25(4)
6 Form 1	SC.S.6.2	13(3)
7 Form 1	SC.S.7.2	4(2)
7 Form 2	SC.S.7.2	2(2)
6 TerraNova	SC.S.6.2	2(2), 5(3), 10(2), 11(2), 13(6), 18(3), 20(2), 22(3), 24(4)
6 TerraNova	SC.S.6.3	7(3)

Alignment of Curriculum Standards and Assessments

Table 3 displays the number of items and points for each assessment form. In the analysis that follows, multiple-point items are given additional weight for alignment purposes. For example, a 3-point item is counted towards the alignment as 3 identically coded 1-point items.

The results of the analysis for each of the four alignment criteria are summarized in Tables 4.1-4.16. More detailed data on each of the criteria are given in Appendix B, in the first three tables. With each table and for each grade, a description of the satisfaction of the alignment criteria for the given grade is provided. The reviewers' debriefing comments provide further detail about the individual reviewers' impressions of the alignment.

In Tables 4.1-4.16, "YES" indicates that an acceptable level was attained between the assessment and the learning goal on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error in the system. "NO" indicates that the criterion was not met by a noticeable margin—10% over an acceptable level for Depth-of-Knowledge Consistency, 10% over an acceptable level for Range-of-Knowledge Correspondence, and .1 under an index value of .7 for Balance of Representation.

Table 3

Number of Items and Point Value by Grade for West Virginia Assessments, Grades 3-9 2008

Grade Level	Number of Items	Number of Multi-Point Items	Total Point Value
3 Form 1	45	0	45
3 Form 2	45	0	45
4 Form 1	45	0	45
4 Form 2	45	0	45
5 Form 1	45	0	45
5 Form 2	45	0	45
6 Form 1	45	0	45
6 Form 2	45	0	45
7 Form 1	45	0	45
7 Form 2	45	0	45
8 Form 1	45	0	45
8 Form 2	45	0	45
9 Form 1	45	0	45
9 Form 2	45	0	45
6 Form TN	25	0	25

Grade 3

The alignment between the WESTEST 2 science grade 3 Form 1 and the West Virginia 21st century standards was acceptable while grade 3 Form 2 and the standards were fully aligned. Both forms had six or more items that corresponded to objectives under each of the three standards. Of the 45 items on each form about 56% (N=23) mapped to objectives under Standard II (Content of Science), 28% (N=13) mapped to objectives under Standard I (Nature of Science), and 16% (N=11) mapped to objectives under Standard III (Application of Science). The DOK levels of the items on the assessment compared favorably to the level of complexity expected by the objectives. Over 70% of the items on each of the forms had a DOK level that was the same or higher than the DOK level of the assigned objective. Whereas the Range of Knowledge Correspondence criterion had an acceptable level in the comparison of Form 2 with all three standards, the acceptable level was only weakly met for two of the standards by Form 1. Reviewers only coded items to 48% of the 11 objectives under each of Standards I and III. Balance of Representation had an acceptable level for all three standards by both forms.

Overall, full alignment between Form 1 and the three standards could be attained by replacing or adding two items, one targeting an objective under Standard I not currently assessed and one targeting an objective under Standard III not currently assessed. Since grade 3 Form 2 was found to be fully aligned with the standards no changes are needed to meet the minimum requirements for alignment used by this

analysis. Reviewers judged that the alignment for Form 1 was acceptable, but did acknowledge there were areas for improvement. One reviewer commented on grade 3 Form 1:

According to my perspective, I found only about half of the objectives to be covered for both standard 1 and 3; for standard 2, I found multiple examples for multiple objectives, with about 1/5 not addressed. From Standard 1: Scientific careers and discoveries were not addressed; supporting statements with facts and use of variables were not addressed. Generally the concepts in Standard 2 are addressed, except for volcanoes, earthquakes, geog features. For standard 3, those missing objectives may be more difficult for question design.

With regards to Form 2, one reviewer indicated that some of the items were really reading-for-inferences items and did not require students to have knowledge of science. The science that students could infer from reading the prompt did map to underlying objectives. Even then the alignment was acceptable for both grade 3 forms, some of the items could be improved to elicit from students their knowledge of standards.

Table 4.1

Summary of Acceptable Levels on Alignment Criteria for Science Grade 3, Form 1, Standards and Assessments for West Virginia Alignment Analysis 2008

Grade 3, Form 1	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.3.I - NATURE OF SCIENCE	YES	YES	WEAK	YES
SC.S.3.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.3.III - APPLICATION OF SCIENCE	YES	YES	WEAK	YES

Table 4.2

Summary of Acceptable Levels on Alignment Criteria for Science Grade 3, Form 2, Standards and Assessments for West Virginia Alignment Analysis 2008

Grade 3 Form 2	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.3.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.3.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.3.III - APPLICATION OF SCIENCE	YES	YES	YES	YES

Grade 4

The alignment between the two grade 4 science WESTEST2 forms and the West Virginia 21st century standards was acceptable for Form 1 and full for Form 2. The 45 items were distributed among the three standards with 24% of the items mapping to Standard I, 59% of the items mapping to Standard II, and 18% of the items mapping to Standard III. This distribution was well over six items for each of the standards. At least 60% of the items on each form mapping to each standard had a DOK level that was the same or higher than the DOK level of the corresponding objective. Range was acceptable for all but Standard I on Form 1. The Form 1 items that targeted objectives under Standard I only corresponded to 45% of the objectives under Standard I, fewer than the 50% needed for an acceptable level. Balance was good for each of the grade 4 assessment forms.

Overall, the alignment between the grade 4 standard and Form 1 was acceptable with only two items needed to be replaced or added to attain full alignment. Form 2 and the standards were found to be fully aligned. Reviewers did not make too many specific comments on the grade 4 assessment forms. One reviewer did note that some items did not precisely match the full intent of the expectation as expressed in the objective. In some of these cases, the refinement of the objective statement would increase the precision of the alignment. This reviewer wrote:

In some cases the standards are so specific that they don't include the broader perspectives on content that are assessed on the test. E.g. objective 4.2.13 specifies differentiating b/tw changes in state but that doesn't exactly match an assessment item that asks a student to know the definition of a gas/solid/liquid. That is just an example, but we had several conversations as a group where folks expressed that they saw parts of the content in an objective, but that the way that it was presented in the objective was so different from the way that it was presented on the assessment that they didn't want to make the match.

Table 4.3
Summary of Acceptable Levels on Alignment Criteria for Science Grade 4, Form 1, Standards and Assessments for West Virginia Alignment Analysis 2008

Grade 4 Form 1	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.4.I - NATURE OF SCIENCE	YES	YES	WEAK	YES
SC.S.4.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.4.III - APPLICATION OF SCIENCE	YES	YES	YES	YES

Table 4.4

Summary of Acceptable Levels on Alignment Criteria for Science Grade 4 Form 2, Standards and Assessments for West Virginia Alignment Analysis 2008

<i>Grade 4 Form 2</i>	<i>Alignment Criteria</i>			
	<i>Categorical Concurrency</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.4.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.4.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.4.III - APPLICATION OF SCIENCE	YES	YES	YES	YES

Grade 5

The alignment between both grade 5 science WESTEST2 forms and the West Virginia science standards was nearly fully aligned. All of the reviewers could only agree on five items that corresponded to objectives under Standard III (Application of Science), one fewer than needed to have an acceptable level on the Categorical Concurrency criterion. The reviewers found on each grade 5 form about 14 items that coded to objectives under Standard I and about 25 items that coded to objectives under Standard II, well over the six items used as the acceptable level. All of the other three criteria were acceptably met for each standard on each of the two forms. Thus, the complexity of the assessment items matched favorably to the complexity of the assigned objectives. The two grade 5 assessment forms had items that targeted over 60% of the objectives underlying the three standards and had an acceptable level for range. The balance was also acceptably met for each assessment form and all three standards.

Overall, the alignment between both grade 5 assessment forms and the standards was acceptable. Only one item would need to be replaced or added to attain full alignment. Even though the data indicate alignment at grade 5, reviewers' comments denoted some issues with the grade 5 assessment items. Compared to the assessments from the other grades analyzed by the grades 3-6 group, reviewers indicated that the fit of items was the worst for grade 5 when compared to grades 3, 4, and 6. Reviewers felt the DOK levels of the grade 5 items were also lower than for other grades and that more items needed more work. So even though the alignment appears to be good, the assessment items and their match to the standards could be improved. Debriefing comments by some of the reviewers noted some of the issues they came across:

- The 5th grade items are poorly written compared with the 3, 4, and 6 grade items. The coverage is not as good of a fit to the objectives.
- I remain concerned that so many questions, even though matched to standards, are actually reading (inferences, usually) questions and can be answered without requiring science knowledge.
- This is the most confusing of the tests that I have been evaluating. I believe that it requires more and better (and correct) science content and

wording in far more items than any before.

Table 4.5
Summary of Acceptable Levels on Alignment Criteria for Science Grade 5, Form 1 Standards and Assessments for West Virginia Alignment Analysis 2008

<i>Grade 5 Form 1</i>	<i>Alignment Criteria</i>			
<i>Standards</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.5.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.5.III - APPLICATION OF SCIENCE	NO (5.5)	YES	YES	YES

Table 4.6
Summary of Acceptable Levels on Alignment Criteria for Science Grade 5, Form 2 Standards and Assessments for West Virginia Alignment Analysis 2008

<i>Grade 5, Form 2</i>	<i>Alignment Criteria</i>			
<i>Standards</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.5.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.5.III - APPLICATION OF SCIENCE	NO (5.83)	YES	YES	YES

Grade 6

Although the alignment between the grade 6 WESTEST2 science assessment forms and the West Virginia 21st science standards was acceptable based on the number of items needed to be replaced or added, both assessment forms were lacking in the measurement of students' knowledge of Standard III (Application of Science). All reviewers only agreed that grade 6 Form 1 had two items and Form 2 had one item that corresponded to one of the six objectives under Standard III. The assessments' coverage of Standard III was noticeably lower than the assessments' coverage for the lower grades. The items did have an appropriate DOK level. Range was good for Standards I and II and balance was appropriate for all three standards with a sufficient number of items to be considered assessed. The grade 3-6 group for science independently coded the grade 6 Form 1 and attained nearly the same results. The only difference was the grade 3-6 group assigned two items to only 31% of the objectives under Standard II whereas the grade 6-9 group assigned two items to 41% of the objectives under Standard II.

Overall, the standards with the grade 6 Form 1 had acceptable alignment whereas the grade 6 Form 2 needed some slight improvement. Four items for Form 1 would need to be added or replaced to have six items that targeted objectives under Standard III and five items would need to be added or replaced to have six items on Form 2 for Standard III. Reviewers were complementary of the grade 6 items. One reviewer did make note regarding some graphing items on Form 1. This reviewer thought some of the content graphed was more appropriate for the senior high school level.

Table 4.7

Summary of Acceptable Levels on Alignment Criteria for Science Grade 6, Form 1 Standards and Assessments for West Virginia Alignment Analysis 2008 (Grade 6-9 Group)

Grade 6, Form 16-9 Group	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.6.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.6.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.6.III - APPLICATION OF SCIENCE	NO (2.83)	YES	WEAK	YES

Table 4.8

Summary of Acceptable Levels on Alignment Criteria for Science Grade 6, Form 1 Standards and Assessments for West Virginia Alignment Analysis 2008 (Grade 3-6 Group)

Grade 6, Form 1 3-6 Group	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.6.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.6.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.6.III - APPLICATION OF SCIENCE	NO (2.83)	YES	NO	YES

Table 4.9

Summary of Acceptable Levels on Alignment Criteria for Science Grade 6, Form 2 Standards and Assessments for West Virginia Alignment Analysis 2008

Grade 6 Form 2	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.6.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.6.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.6.III - APPLICATION OF SCIENCE	NO (1.33)	NT	NT	NT

Grade 7

As for the other grades, the alignment between the two grade 7 WESTEST2 science assessment forms and the grade 7 West Virginia 21st century standards was acceptable. Both assessment forms only had three items that targeted objectives under Standard III (Application of Science). From 63% to 68% of the 45 items mapped to Standard II and about 30% of the items mapped to Standard I, well over the minimum of six items needed for an acceptable level for the Categorical Concurrence criterion. The Depth-of-Knowledge Consistency criterion was acceptably met for all three standards on each of the two forms. At least 69% of the items on each form had a DOK level that was comparable to the DOK level of the assigned objective. Range was good for Standards I and II on each form, over 60% of the objectives with at least one item. However, the low number of items corresponding to Standard III contributed to a low range for this standard. The items were evenly distributed among the objectives under all of the objectives to have an acceptable balance.

Table 4.10

Summary of Acceptable Levels on Alignment Criteria for Science Grade 7, Form 1 Standards and Assessments for West Virginia Alignment Analysis 2008

Grade 7 Form 1	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.7.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.7.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.7.III - APPLICATION OF SCIENCE	NO (3.33)	YES	WEAK	YES

Table 4.11

Summary of Acceptable Levels on Alignment Criteria for Science Grade 7, Form 2 Standards and Assessments for West Virginia Alignment Analysis 2008

<i>Grade 7 Form 2</i>	<i>Alignment Criteria</i>			
<i>Standards</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.7.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.7.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.7.III - APPLICATION OF SCIENCE	NO (3.5)	YES	WEAK	YES

Overall, the alignment between the two grade 7 assessment forms and standards was acceptable. For each of the two forms three items would need to be replaced or added to attain full alignment. The new items would need to correspond to objectives under Standard III, at least one that is not currently assessed. In general, the reviewers felt the grade 7 items were appropriate. One reviewer did note that reviewers had a difficulty coming to agreement on the correct answer to some items on the grade 7 assessment. Their comments (Appendix C) should be reviewed for specific issues on individual items.

Grade 8

The alignment between the two grade 8 science WESTEST2 assessment forms and the West Virginia 21st century standards was acceptable. The main alignment issue was that the two assessment forms had too few items that targeted objectives under Standard III as for the prior grades. The reviewers agreed that the grade 8 Form 1 had three items (Items 34, 42, and 43) that targeted objectives under Standard III. Grade 8 Form 2 had five items (Items 6, 27, 37, 39, and 42). The two grade 8 assessment forms were fully aligned with Standards I and II. For each of these two standards the assessments had 11 or more items that corresponded to each standard. Over 58% of the items coded to these two standards had a DOK level that was the same or higher than the DOK level of the assigned objective. These items targeted more than 60% of the underlying objectives to have an acceptable range and were evenly distributed among the objectives to have an acceptable balance.

Overall, the alignment for grade 8 science was acceptable for the two WESTEST2 forms with only three items for Form 1 and one item for Form 2 needed to be replaced or added to attain full alignment. These additional items would need to target objectives under Standard III. Some care would be needed to select items for Form 1 so that the items had an appropriate DOK level and targeted at least two objectives not currently assessed. The items that targeted Standard III on both forms would be sufficient to have one assessment form aligned with that standard. For example, Form 1 Items 42 and 43 along with Form 2 Items 6, 27, 37, and 39 would be fully aligned considering all four alignment criteria. The alignment would be stronger if one of the items had a DOK level

3 that all reviewers could agree upon, similar in complexity to Item 43 on Form 1 (two reviewers judged had a DOK level 3) and Item 39 on Form 2 (three reviewers judged had a DOK level 3). Reviewers' comments supported that the alignment was acceptable. They did make some comments about specific items (see Appendix C).

Table 4.12

Summary of Acceptable Levels on Alignment Criteria for Science Grade 8, Form 1 Standards and Assessments for West Virginia Alignment Analysis 2008

Grade 8 Form 1	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.8.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.8.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.8.III - APPLICATION OF SCIENCE	NO (3.67)	WEAK	WEAK	YES

Table 4.13

Summary of Acceptable Levels on Alignment Criteria for Science Grade 8, Form 2 Standards and Assessments for West Virginia Alignment Analysis 2008

Grade 8 Form 2	Alignment Criteria			
Standards	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.8.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.S.8.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.8.III - APPLICATION OF SCIENCE	NO (1.0)	YES	YES	YES

Grade 9

The alignment between the two forms of the WESTEST2 grade 9 physical science assessments and the West Virginia 21st century standards was acceptable. As for the previous grades, reviewers judged the two forms each had fewer than six items that targeted objectives under Standard III (Application of Science). Form 1 had, on the average, 4.5 items while Form 2 had 5.33 items that were mapped to Standard III. In addition, the items that mapped to Standard III on both forms were slightly lower in DOK when compared to the DOK levels of the matched objectives and did not target a sufficient percentage of the objectives. Form 2 also had a DOK weakness for Standard I. Both assessment forms were fully aligned with Standard II. Standard I and Form 1 were also found to be fully aligned.

Table 4.14

Summary of Acceptable Levels on Alignment Criteria for Science Grade 9, Form 1 Standards and Assessments for West Virginia Alignment Analysis 2008

<i>Grade 9, Form 1</i>	<i>Alignment Criteria</i>			
<i>Standards</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.9.I - NATURE OF SCIENCE	YES	YES	YES	YES
SC.9.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.9.III - APPLICATION OF SCIENCE	NO (4.5)	WEAK	WEAK	YES

Table 4.15

Summary of Acceptable Levels on Alignment Criteria for Science Grade 9, Form 2 Standards and Assessments for West Virginia Alignment Analysis 2008

<i>Grade 9, Form 2</i>	<i>Alignment Criteria</i>			
<i>Standards</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.9.I - NATURE OF SCIENCE	YES	WEAK	YES	YES
SC.9.II - CONTENT OF SCIENCE	YES	YES	YES	YES
SC.S.9.III - APPLICATION OF SCIENCE	NO (5.33)	WEAK	WEAK	YES

Overall, the alignment for grade 9 physical science was acceptable with only two items needed to be replaced or added to both forms in order to attain full alignment. Two additional items for Form 1 would need to be mapped to Standard III to have full alignment. These two items would have to target objectives not currently assessed and with a DOK level that is the same or higher than the assigned objective. For Form 2, one item would need to be added that targets an additional objective under Standard III and with an appropriate DOK level. Also, one current item that corresponds to an objective under Standard I would need to be replaced with an item that has an appropriate DOK level. By combining items from the two grade 9 assessment forms, it would be possible to have one form that would be fully aligned. For example, Items 12, 32, 33, and 34 from Form 1 and Items 10 and 28 from Form 2 would satisfy the minimum levels for the four alignment criteria. All but one reviewer judged that Item 34 on Form 1 had a DOK level 3, the item with the highest level of complexity. The alignment to Standard III would be stronger if one or two other items mapping to Objectives 2, 4 and 6 had a DOK level 3. Reviewers in general indicated the alignment was acceptable and only made comments on a few individual items on the grade 9 physical science assessments.

TerraNova Grade 6

The alignment between the West Virginia 21st century grade 6 science standards and the TerraNova grade 6 form needed major improvement. Of the seven TerraNova forms, one for each grade, time was available to only analyze the form for grade 6. Reviewers found the TerraNova grade 6 science form had a sufficient number of items for Standard I (N=8) and Standard II (N=14), but was lacking in items for Standard III (N=1). Only about one-third of the items that mapped to objectives under Standard II had a DOK level that was the same or higher than the DOK level of the assigned objective. The TerraNova assessment form also did not have a sufficient breadth in content coverage. The range criterion was not acceptable for any of the three standards. However, balance was good for all three standards.

Overall, the TerraNova grade 6 form was found to need major improvement to be aligned with the grade 6 science standards. At least 13 items would need to be added or replaced to have full alignment. Two items of the eight items that targeted objectives under Standard I would need to be replaced by items that target two objectives not currently assessed. For Standard II, six items with an appropriate DOK level would need to be replaced to target six additional objectives. Five items would need to be added that target objectives under Standard III. These items are needed to increase the total number of items for Standard III to six and would need to assess at least two or three objectives that are not currently targeted by existing items. Reviewers' comments also noted alignment issues between the grade 6 science standards and the TerraNova assessment. The large number of items coded to the generic objective under Standard II by two or more reviewers (Table 2) stresses the lack of close fit between the TerraNova assessment and the West Virginia standards. One reviewer's summary represented the view by other reviewers:

It seems that many WV important conceptual ideas are not addressed in this assessment. It also seems that some of the items in the TN battery are based on important concepts that are not specifically addressed in the WV standards/objectives.

Table 4.16

Summary of Acceptable Levels on Alignment Criteria for Science Grade 6, Form TN, Standards and Assessments for West Virginia Alignment Analysis 2008

<i>Grade 6, Form TN</i>	<i>Alignment Criteria</i>			
	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
SC.S.6.I - NATURE OF SCIENCE	YES	YES	WEAK	YES
SC.S.6.II - CONTENT OF SCIENCE	YES	NO	NO	YES
SC.S.6.III - APPLICATION OF SCIENCE	NO (1.67)	YES	NO	YES

Source of Challenge Issue and Reviewers' Comments

Reviewers were instructed to document any source-of-challenge issue and to provide any other comments they may have. These comments can be found in Tables (grade).5 and (grade).7 in Appendices C and F. After coding each grade-level assessment, reviewers also were asked to respond to five debriefing questions. All of the comments made by the reviewers are given in Appendices D and G. The notes in general offer an opinion on the item or give an explanation of the reviewers' coding.

Reliability Among Reviewers

The overall intraclass correlation among the science reviewers' assignment of DOK levels to items was high for six reviewers for grades 3-9 (Table 5). An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers. The intraclass correlations for all of the analysis were higher than 0.8. A pairwise comparison was used to determine the degree of reliability of reviewer coding at the objective and standard levels. The standard pairwise comparison values (all above 0.85) and the objective pairwise comparison values (above 0.65 except for the TerraNova analysis) were greater than for most alignment studies and indicate reasonable agreement among reviewers. The agreement values, however, were not computed using independent judgments, but were computed after reviewers adjudicated their codings. After reviewers discussed the codings of items where there was a large variance among the reviewers, reviewers could change their personal codings if they felt compelled to do so.

Table 5

Intraclass and Pairwise Comparisons, West Virginia Alignment Analysis for Science Grades 3-9 Assessments

Grade	Intraclass Correlation	Pairwise Comparison:	Pairwise: Objective	Pairwise: Standard
3 Form 1	.88	.71	.80	.92
3 Form 2	.93	.48	.74	.87
4 Form 1	.89	.74	.78	.93
4 Form 2	.88	.76	.86	.98
5 Form 1	.85	.71	.78	.90
5 Form 2	.84	.66	.74	.88
6 Form 1	.82	.62	.70	.86
6 Form 2	.86	.70	.66	.86
7 Form 1	.91	.74	.68	.86
7 Form 2	.91	.78	.84	.94
8 Form 1	.90	.78	.77	.90
8 Form 2	.94	.80	.81	.92
9 Form 1	.90	.75	.82	.90
9 Form 2	.82	.76	.74	.90
6 Form TN	.92	.82	.55	.90

Summary

A three day alignment institute was held September 17-19, 2008 in Charleston, West Virginia to analyze the pre-field test forms of the WESTEST2 and TerraNova with the West Virginia 21st century science standards. Two groups of six reviewers each participated in the institute. One group analyzed assessments and standards for grades 3-6 and one group analyzed these documents for grades 6-8 and grade 9 physical science. Both groups independently analyzed the alignment between the standards and grade 6 Form 1. Half of the reviewers were from West Virginia and half were from other states. The reviewers included science education content experts, state science supervisors, and science teachers. Two forms of the WESTEST2 assessment for each grade and one TerraNova grade six form were analyzed.

For nearly all of the fifteen science forms analyzed, the alignment between the WESTEST2 assessments and the West Virginia 21st century science standards was acceptable. The science standards and two WESTEST2 forms were found to be fully aligned (grade 3 Form 2 and grade 4 Form 2). For the other WESTEST2 science forms some minor alignment issue was detected. All of the WESTEST2 forms for grades 5 through 9 had fewer than six items that targeted objectives under Standard III (Application of Science). The lower number of items for Standard III was generally accompanied by too many items with too low of a DOK level or a lack of range. The TerraNova grade 6 assessment and the grade 6 standards needed major improvement.

Reviewers commented that the science items were generally reasonable. The one repeated comment by reviewers in the grade 3-6 group was that some of the items were more of reading items where the students were required to infer from the prompt the science rather than recall or apply conceptual knowledge. More specific comments on individual items are included in the appendices. Overall, the alignment for science was at least acceptable with fewer than five items needed to be replaced or added to attain full alignment as summarized in the table below.

Summary Table

Percent of West Virginia Mathematics Standards with Acceptable Level on Each Alignment Criteria for Grade 3-8 and 9 Physical Science for WESTEST2 Analysis

Grade	<i>Categorical Concurrence</i> (six or more items)	<i>Depth-of-Knowledge Consistency</i> (50% at/above)	<i>Range of Knowledge</i> (50% of objectives)	<i>Balance of Representation</i> (without possible weakness)	<i>Estimated Range of Items per to be Added or Replaced for Full Alignment</i>
3 Form 1	100	100	33	100	2
3 Form 2	100	100	100	100	0
4 Form 1	100	100	67	100	2
4 Form 2	100	100	100	100	0
5 Form 1	67	100	100	100	1
5 Form 2	67	100	100	100	1
6 Form 1	67	100	67	100	4

6 Form 2	67	67	67	67	5
7 Form 1	67	100	100	100	3
7 Form 2	67	100	100	100	3
8 Form 1	67	67	67	100	3
8 Form 2	67	100	100	100	1
9 Form 1	67	67	67	100	2
9 Form 2	67	33	67	100	2
6 TerraNova	67	67	0	100	13

Categorical Concurrence	>6 items
Depth-of-Knowledge	>50% with DOK level the same or higher than level of corresponding Objectives
Range-of-Knowledge	>70% of objectives under a standard
Balance of Representation	A possible weakness if one or more objectives with a relative large number of items (e.g. five or more than the objective with the next highest number of items)

References

- Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25(1), 47-55.
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.