

Missouri Assessment Program End-of-Course (EOC) Assessment Forms Alignment Validation Study: Technical Report

**Leslie R. Taylor
Hilary L. Campbell
Richard C. Deatz
Rebecca N. Dvorak
Lisa E. Koger
Arthur A. Thacker**

Prepared for: Missouri Department of Elementary and Secondary Education
205 Jefferson Street
P.O. Box 480
Jefferson City, Missouri 65102

Prepared under: Contract No: C308004001-003
March 8, 2011

Missouri Assessment Program End-of-Course (EOC) Assessment Forms Alignment Validation Study: Technical Report

**Leslie R. Taylor
Hilary L. Campbell
Richard C. Deatz
Rebecca N. Dvorak
Lisa E. Koger
Arthur A. Thacker**

Prepared for: Missouri Department of Elementary and Secondary Education
205 Jefferson Street
P.O. Box 480
Jefferson City, Missouri 65102

Prepared under: Contract No: C308004001-003
March 8, 2011

EXECUTIVE SUMMARY

Scope of Work

The Missouri Department of Elementary and Secondary Education (DESE) requested an external independent alignment study of the Missouri Assessment Program End-of-Course (EOC) Assessments for English II, Algebra I, and Biology. Specifically, the study evaluated the alignment of a single form from each of the Fall 2009, Summer 2010, and Spring 2011 assessments to the Missouri Course-Level Expectations (CLEs)¹. Missouri uses the EOC tests in federal and state accountability programs. DESE awarded the Human Resources Research Organization (HumRRO) the contract to conduct this alignment study.

DESE requested the alignment study in order to meet both state and federal requirements. The federal requirements of the U.S. Department of Education (USDE) stem from the *No Child Left Behind Act* (NCLB) of 2001. NCLB challenges each state to establish a coherent assessment system based on solid academic standards. This law calls for states to provide independent evidence of the validity of their assessments used to calculate Adequate Yearly Progress (AYP). All states receiving Title I funds must present evidence of establishing a fair and consistent assessment system based on rigorous standards, sufficient alignment between standards and assessments, and high-quality educational results.

An alignment review can provide one form of evidence supporting the validity of a state's assessment system. Alignment results should demonstrate that the assessments represent the full range of content standards and that the assessments measure student knowledge in the same manner and at the same level of complexity as specified in the content standards. All aspects of the state assessment system must coincide, including the academic content standards, achievement standards (linked to cut scores), performance level descriptors, and each assessment.

Methodology

HumRRO convened three content panels of Missouri educators to review the EOC test forms. These panelists included current teachers, administrators, and curriculum specialists or district coordinators.

HumRRO conducted the reviews in Jefferson City, Missouri, on November 2 and 3, 2010, using the Webb alignment method (1997; 1999; 2005). As part of this method, reviewers rate individual test items on cognitive complexity and content assessed relative to state content standards. The Webb procedure for evaluating alignment of the assessment to the content standards involves analysis of four alignment measures. These measures indicate how well an assessment covers the content standards in terms of content breadth and depth. The four alignment indicators include:

¹ Missouri Course-Level Expectations can be found at <http://dese.mo.gov/divimprove/curriculum/GLE>

- (1) Categorical concurrence – determines the degree of overall content coverage by the assessment for each content strand.
- (2) Range-of-knowledge representation – indicates the specific content expectations (e.g., standard, course-level expectations) assessed within each strand.
- (3) Balance-of-knowledge representation – provides a statistical index reflecting the distribution of assessed content within each strand (i.e., how evenly the content is assessed.)
- (4) Depth-of-knowledge (DOK) consistency – compares the cognitive complexity ratings of the items with the complexity ratings of each content standard.

Summary of Results

The extent of alignment to the Missouri CLEs varied per content area and test form.

Key Findings and Conclusions

English II. Regarding English II, all of the English II test forms covered the breadth of the Reading CLEs quite well. The test forms reviewed for each course also exhibited some gaps that DESE may wish to review to improve alignment. Regarding English II, we point to two potential alignment issues. First, the depth-of-knowledge assessed does not match the CLEs for many items targeting Reading-Nonfiction. Second, the test forms assess *Writing* content in a narrow way, an outcome that can be explained in part by the design of the test blueprint and State suspension of performance events (Session II). Assessing writing skills on a state-level assessment can be challenging, particularly if no writing component is in place as is the case currently. As a result, reviewers determined that the selected-response portion of the assessment covers a single Writing CLE (Text Development - Conventions of English) with approximately five items. With the performance event (included in the Fall 2009 administration), reviewers still matched items to only two CLEs total (Conventions of English and Forms/Types of Writing). Thus, the test forms reviewed as part of this study partly align to the content standards.

Algebra I. The Algebra I results suggest that the test forms fully align to the breadth of the CLEs across strands. Some test items on each form may require review of DOK consistency relative to the Data and Probability strand. Reviewers rated over half of items on the 2010 and 2011 forms as below the DOK level of the targeted CLEs. In most cases, the degree of discrepancy involved an adjacent mismatch (i.e., item DOK=2; CLE DOK=3). The 2009 test form did surpass the minimum criterion (M=58% of items matched the DOK of the CLEs).

Biology. For Biology, all test forms exhibited full alignment on content breadth and depth relative to the Living Organisms and Ecology strands. However, the Biology test forms show gaps specifically in assessment of Scientific Inquiry in breadth and

depth. The limited assessment of the breadth of this strand overall does correspond with the design of Missouri CLEs for Biology, however. The Scientific Inquiry strand is intended for assessment only by performance events because it requires students to demonstrate integrated understanding of scientific principles, particularly the application of experimental procedures. Given that the State was forced to eliminate performance events from the 2010 and 2011 Biology test forms, this circumstance makes it challenging for DESE to assess this content at this time. Furthermore, we recognize that these content expectations serve as a process strand to be assessed along with other science content strands. DESE will want to determine how to handle this strand as they transition to a new test vendor. We do note that, for the 2009 test form with performance events, the depth of content assessed did not match the CLEs under Scientific Inquiry for some performance events. Analyses of the 2009 form with performance events show that the majority of performance items assess students at a lower level of cognitive complexity than expected (i.e., item DOK=2 and CLE DOK=3). Thus, if DESE can pursue state-level assessment of Scientific Inquiry in the future, an increase in cognitive complexity may be needed.

Alignment of EOC Test Forms to Missouri Course-Level Expectations

Tables 1 and 2 provide summary conclusions on the alignment of the EOC test forms reviewed relative to the Missouri CLEs for English II, Algebra I, and Biology. The conclusions are based on the following decision criteria (Webb, 2005):

- Fully aligned – assessments align to all content strands (100%);
- Highly aligned – assessments align to the majority of strands (70%–99%);
- Partially aligned – assessments align well to some strands (50%–69%); and
- Weakly aligned – assessments align to less than half the strands (below 50%).

The conclusions in Table 1 focus on the alignment analyses on the 2009 test forms for each course, including multiple-choice and performance event items. In comparison, Table 2 only displays results for analyses on the multiple-choice items.

Table 1 Summary Alignment Conclusions for 2009 Test Form per Course by Webb Alignment Indicator (Multiple-Choice and Performance Event Items)

Test Forms	Alignment Conclusions per Webb Indicator			
	Categorical Concurrency	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
English II	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (67%)	Highly aligned (80%)
Algebra I	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)
Biology	Fully aligned (100%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)

Table 2 Summary Alignment Conclusions per Course Test Form by Webb Alignment Indicator (Multiple-Choice Items Only)

Test Forms	Alignment Conclusions per Webb Indicator											
	Categorical Concurrence			Depth-of-Knowledge Consistency			Range-of-Knowledge Correspondence			Balance-of-Knowledge Representation		
	Fall 2009	Summer 2010	Spring 2011	Fall 2009	Summer 2010	Spring 2011	Fall 2009	Summer 2010	Spring 2011	Fall 2009	Summer 2010	Spring 2011
English II	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Weakly aligned (37%)	Weakly aligned (37%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)
Algebra I	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Partially aligned (67%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)
Biology	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)

Recommendations

English II

1. **Review DOK assessed by Reading items relative to the Missouri CLEs (DOK consistency).** Panelists' ratings of item DOK for each test form indicate that some items assess student knowledge below the DOK level of corresponding CLEs. This was particularly true for 2010 and 2011 items targeting Nonfiction. Changing DOK for as few as three items covering Reading would improve alignment above the minimum criterion.
2. **Review the content distribution of items assessing Writing CLEs.** While the test blueprint does specify that many Writing CLEs should be assessed by performance events, DESE may wish to determine if writing could be assessed in a more even manner. If suspension of performance events continues to be a necessity, one approach in future administrations for increasing alignment could be to include additional selected-response format items requiring text evaluation for writing structure. We recognize that this solution may involve item development, which may be cost prohibitive in the immediate future. A second option may be to tie course-level student writing products to the state assessment. A few states have pursued this option by including writing components graded by teachers, based on a state-developed rubric, as part of scores. Finally, DESE should (and probably has already) emphasized to teachers that course-level assessment of writing is critical to ensure sufficient coverage of these skills.

Algebra I

1. **Review DOK for some items on 2010 and 2011 test forms relative to the Data and Probability strand (DOK consistency).** Reviewers' ratings indicate a lower level cognitive complexity for items assessing the Data and Probability strand, particularly for the 2010 and 2011 test forms. As noted for English II, improving DOK alignment could involve minor item edits to stems and/or response options for one to three items. However, we recognize that discrepancy between items and CLEs for this strand is inevitable given that these particular CLEs, and Data and Probability content in general, require students to demonstrate knowledge at a higher level of processing.

Biology

1. **Review the test forms for coverage of the Scientific Inquiry strand (all Webb indicators).** While the test forms very align well to the Living Organisms and Ecology strands, reviewers found substantial gaps in assessment of Scientific Inquiry. However, this issue is not unexpected, as noted in the earlier discussion. First, this strand is intended for assessment only by performance events because it requires students to demonstrate integrated understanding of scientific principles, particularly the application of

experimental procedures. Thus, the limited assessment of Scientific Inquiry, even on the 2009 test form including performance events, does accurately reflect the intention of the standards and the test blueprint. Second, and consequently, the elimination of performance events from Biology test forms makes it challenging for the state to assess this content at this time.

If the circumstances for DESE change and additional funds become available, we offer several recommendations for item development. In a similar way as recommended for the Writing strand for English II, we suggest considering whether some CLEs under this strand could be assessed by selected-response items. As an embedded strand, selected-response items could address scientific inquiry along with a primary content strand, which may be possible with current items. Alternatively, DESE could pursue item development of basic knowledge of the scientific process. While not ideal given that the intention of this strand is to encourage student reasoning and analysis, some representation of this strand would achieve greater breadth of the State-level content expectations.

**MISSOURI ASSESSMENT PROGRAM END-OF-COURSE (EOC) ASSESSMENT
FORMS ALIGNMENT VALIDATION STUDY: TECHNICAL REPORT**

TABLE OF CONTENTS

Chapter 1: Introduction	1
Chapter 2: Alignment Study Design and Methodology	3
Webb Alignment Method.....	3
EOC Alignment Reviews.....	4
<i>Panelists</i>	4
<i>Materials</i>	4
<i>Procedures</i>	5
Statistical Method and Analysis.....	6
<i>Inter-rater Agreement Results</i>	7
<i>Webb Alignment Measures</i>	7
Chapter 3 Results: English II	11
Inter-rater Agreement Results.....	11
Webb Alignment Results.....	11
<i>Categorical Concurrence</i>	12
<i>DOK Consistency</i>	14
<i>Range-of-Knowledge</i>	16
<i>Balance-of-Knowledge Representation</i>	17
Summary and Discussion of Webb Alignment Results for English II.....	18
Chapter 4 Results: Algebra I	21
Inter-rater Agreement Results.....	21
Webb Alignment Results.....	22
<i>Categorical Concurrence</i>	22
<i>DOK Consistency</i>	23
<i>Range-of-Knowledge</i>	24
<i>Balance-of-Knowledge Representation</i>	25
Summary and Discussion of Webb Alignment Results for Algebra I.....	26
Chapter 5 Results: Biology	28
Inter-rater Agreement Results.....	28
Webb Alignment Results.....	29
<i>Categorical Concurrence</i>	29
<i>DOK Consistency</i>	30
<i>Range-of-Knowledge</i>	32
<i>Balance-of-Knowledge Representation</i>	33
Summary and Discussion of Webb Alignment Results for Biology.....	33
Chapter 6: Summary and Recommendations	36
Recommendations.....	39
<i>English II</i>	39

Algebra I 39
Biology 40
References 41

LIST OF TABLES

Table 1 Summary Alignment Conclusions for 2009 Test Form per Course by Webb Alignment Indicator (Multiple-Choice and Performance Event Items).....iv
 Table 2 Summary Alignment Conclusions per Course Test Form by Webb Alignment Indicator (Multiple-Choice Items Only)..... v
 Table 2.1 Demographic Characteristics of EOC Panelists 4
 Table 2.2 Item Composition for Fall 2009, Summer 2010, and Spring 2011 EOC Test Forms 5
 Table 2.3 Webb’s Depth-of-knowledge Rating Scale 8
 Table 3.1 Intraclass Correlation Coefficients on Item DOK Ratings for English II 11
 Table 3.2. Pairwise Comparisons on Reviewer Content Agreement for English II 11
 Table 3.3 Summary of Categorical Concurrence Results for English II 2009 Test Form (Multiple-Choice and Performance Event Items) 13
 Table 3.4 Summary of Categorical Concurrence Results for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only) 13
 Table 3.5 Summary of DOK Consistency by Items for English II 2009 Test Form (Multiple-Choice and Performance Event Items) 14
 Table 3.6 Summary of DOK Consistency by Items for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only) 15
 Table 3.7 Summary of DOK Consistency by CLEs for English II 2009 Test Form (Multiple-Choice and Performance Event Items) 15
 Table 3.8 Summary of DOK Consistency by CLEs for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only) 16
 Table 3.9 Summary of Range-of-Knowledge Results for English II 2009 Test Form (Multiple-Choice and Performance Event Items) 16
 Table 3.10 Summary of Range-of-Knowledge Results for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only) 17
 Table 3.11 Summary of Balance-of-Knowledge Results for English II 2009 Test Form (Multiple-Choice and Performance Event Items) 18
 Table 3.12 Summary of Balance-of-Knowledge Results for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only) 18
 Table 3.13 Summary Alignment Outcomes per Webb Criterion for English II 2009 Test Form with Multiple-Choice and Performance Event Items 19
 Table 3.14 Summary Alignment Outcomes per Webb Criterion for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only) 19
 Table 4.1 Intraclass Correlation Coefficients on Item DOK Ratings for Algebra I 21
 Table 4.2. Pairwise Comparisons on Reviewer Content Agreement for Algebra I 21
 Table 4.3 Summary of Categorical Concurrence Results for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items) 22

Table 4.4 Summary of Categorical Concurrence Results for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	22
Table 4.5 Summary of DOK Consistency by Items per Test Form for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)	23
Table 4.6 Summary of DOK Consistency by Items per Test Form for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	23
Table 4.7 Summary of DOK Consistency by CLEs per Test Form for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)	24
Table 4.8 Summary of DOK Consistency by CLEs per Test Form for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	24
Table 4.9 Summary of Range-of-Knowledge Results for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)	25
Table 4.10 Summary of Range-of-Knowledge Results for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	25
Table 4.11 Summary of Balance-of-Knowledge Results for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items Only)	25
Table 4.12 Summary of Balance-of-Knowledge Results for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	26
Table 4.13 Summary Alignment Outcomes per Webb Criterion for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)	27
Table 4.14 Summary Alignment Outcomes per Webb Criterion for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	27
Table 5.1 Intraclass Correlation Coefficients on Item DOK Ratings for Biology	28
Table 5.2. Pairwise Comparisons on Reviewer Content Agreement for Biology	29
Table 5.3 Summary of Categorical Concurrence Results for Biology 2009 Test Forms (Multiple-Choice and Performance Event Items)	29
Table 5.4 Summary of Categorical Concurrence Results for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	30
Table 5.5 Summary of DOK Consistency by Items per Test Form for Biology 2009 Test Forms (Multiple-Choice and Performance Event Items)	30
Table 5.6 Summary of DOK Consistency by Items per Test Form for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	31
Table 5.7 Summary of DOK Consistency by CLEs per Test Form for Biology 2009 Test Forms (Multiple-Choice and Performance Event Items)	31
Table 5.8 Summary of DOK Consistency by CLEs per Test Form for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	32
Table 5.9 Summary of Range-of-Knowledge Results for Biology 2009 Test Form (Multiple-Choice and Performance Event Items)	32
Table 5.10 Summary of Range-of-Knowledge Results for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	32
Table 5.11 Summary of Balance-of-Knowledge Results for Biology 2009 Test Form (Multiple-Choice and Performance Event Items)	33
Table 5.12 Summary of Balance-of-Knowledge Results for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)	33
Table 5.13 Summary Alignment Outcomes per Webb Criterion for Biology 2009 Test Form (Multiple-Choice and Performance Event Items)	34

Table 5.14 Summary Alignment Outcomes per Webb Criterion for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only) 35

Table 6.1. Summary Alignment Conclusions for 2009 Test Form per Course by Webb Alignment Indicator (Multiple-Choice and Performance Event Items)..... 37

Table 6.2. Summary Alignment Conclusions per Course Test Form by Webb Alignment Indicator (Multiple-Choice Items Only)..... 38

MISSOURI ASSESSMENT PROGRAM END-OF-COURSE (EOC) ASSESSMENT FORMS ALIGNMENT VALIDATION STUDY: TECHNICAL REPORT

Chapter 1: Introduction

The Missouri Department of Elementary and Secondary Education (DESE) requested an external independent alignment study of the Missouri Assessment Program End-of-Course (EOC) assessments for English II, Algebra I, and Biology. Specifically, the study evaluated the alignment of a single form from each of the Fall 2009, Summer 2010, and Spring 2011 assessments to the Missouri Course-Level Expectations (CLEs)². Missouri uses the EOC tests in federal and state accountability programs. DESE awarded the Human Resources Research Organization (HumRRO) the contract to conduct this alignment study.

DESE requested the alignment study in order to meet state and federal requirements. The federal requirements of the U.S. Department of Education (USDE) stem from the *No Child Left Behind Act* (NCLB) of 2001. NCLB challenges each state to establish a coherent assessment system based on solid academic standards. This law calls for states to provide independent evidence of the validity of their assessments used to calculate Adequate Yearly Progress (AYP). All states receiving Title I funds must present evidence of establishing a fair and consistent assessment system based on rigorous standards, sufficient alignment between standards, and assessments and high-quality educational results.

An alignment review can provide one form of evidence supporting the validity of a state's assessment system. The term *alignment* in this context refers to the degree of consistency evident in instruction and measurement of the state's academic content standards. Alignment results should demonstrate that the assessments represent the full range of content standards and measure student knowledge in the same manner and at the same level of complexity as specified in the content standards. All aspects of the state assessment system must coincide, including the academic content standards, achievement standards (linked to cut scores), performance level descriptors, and each assessment. An alignment study can evaluate the strength of any one, or all, of these relationships.

In general, alignment evaluations of any assessment reveal the breadth, or scope, of knowledge as well as the depth-of-knowledge, or cognitive processing, expected of students by the state's content standards. Alignment analyses help answer questions such as the following:

- How much and what type of content is covered by the assessment?
- Is the content in the assessment, or other standards, sufficiently similar to the expectations of the full content standards?
- Are students asked to demonstrate this knowledge at the same level of rigor as expected in the full content standards?

² Missouri Course-Level Expectations can be found at <http://dese.mo.gov/divimprove/curriculum/GLE>

Organization and Contents of the Report

This report contains six chapters. Chapter 2 describes the alignment method and test review details, including panelist characteristics, materials, procedures, and statistical method. Chapters 3 through 5 provide alignment results for each EOC content test (English II, Algebra I, and Biology, respectively). Chapter 6 provides recommendations for DESE to strengthen the alignment of the EOC tests over time.

HumRRO provides additional information in the appendices of this report. Appendices A through C contain tables that provide details on the content alignment results per course assessment. Appendix D includes a summary of panelists' comments on their ratings based on the type of comment provided. Appendix E provides examples of rating forms and training materials used in the alignment workshops.

Chapter 2: Alignment Study Design and Methodology

In this chapter, we discuss key concepts related to alignment research, followed by a description of the alignment evaluations and methods used as part of the Missouri EOC forms validation study.

Webb Alignment Method

Several methods of alignment exist. For the Missouri alignment studies, HumRRO applied the Webb alignment method designed for use with standard large-scale assessments. This method, which has been refined over time (e.g., Webb, 1997; 1999; 2005), is supported by the Council of Chief State School Officers (CCSSO) and has been applied in many states.

The Webb method includes four major criteria to evaluate alignment. These criteria link with statistical procedures used to assess how well individual portions of the assessments and content standards documents actually match. The four alignment criteria are as follows: (a) categorical concurrence, (b) depth-of-knowledge consistency, (c) range-of-knowledge correspondence, and (d) balance-of-knowledge representation.

Categorical concurrence is a basic measure of alignment between content standards and test items. This term refers to the proportion of overlap between the content stated in the standards document and that assessed by items on the test.

Depth-of-knowledge (DOK) refers to the type of cognitive processing required by items and content standards. For example, is a student expected to simply identify or recall basic facts, or is the student expected to use reasoning in manipulating information or strategizing? Using Algebra I as an example, a student may be asked to identify the appropriate use of a decimal among several answer choices. This task should be less complex than trying to explain the concept of a decimal and how and why it can be moved. The purpose of using DOK as a measure of alignment is to determine whether a test item (or performance task) and its corresponding standard are written at the same level of cognitive complexity. Reviewers make two separate judgments about cognitive complexity, one for the standard and one for the item. These two judgments are compared to determine whether the item is written at the same level as the standard. Webb refers to this comparison as *depth-of-knowledge consistency*.

The ***range-of-knowledge correspondence*** measure analyzes the breadth of knowledge represented by test items in more detail. Categorical concurrence simply notes whether a sufficient number of items on the test covers each general content topic (individual strands). However, states usually outline more specific *content objectives*, or standards, under each strand. The range indicates the number of standards assessed by items.

Finally, **balance-of-knowledge representation** examines the distribution of the content assessed by items. This analysis focuses on whether content emphasis on the assessment is comparable to the state standards document. Content balance is determined by calculating an index, or score, based on the distribution of items assessing each strand. Each strand should meet or surpass a minimum index to demonstrate adequate balance.

EOC Alignment Reviews

In this section, we describe the 2010 EOC alignment workshops, including panelists, materials, and procedures.

Panelists

HumRRO convened three content panels of Missouri educators to review the EOC test forms. All panelists who served on panels are current teachers with expertise in the content area they reviewed.

To establish review panels, HumRRO received district contact information from DESE, sent inquiries of interest to districts and individuals across the state, and accepted applications to serve on the panels. HumRRO developed a pool of highly qualified candidates based on applications. From this pool, HumRRO selected panels, with a target of five panelists per course, by considering several factors in an effort to balance panels appropriately: (a) region of origin in Missouri, (b) other demographic factors (e.g., rural/suburban, gender), and (c) status as a new or former panelist. Table 2.1 presents the characteristics of the panelists per EOC content area.

Table 2.1 Demographic Characteristics of EOC Panelists

Content Panel	Group Size	Panelist Status		Gender		Number of Panelists per Region								
		New panelist	Previous panelist	M	F	S E	Heart of MO	KC	NE	NW	SC	SW	STL	Central
English II	5	2	3	0	5	0	1	1	1	1	0	1	0	0
Algebra I	4	2	2	0	4	1	1	0	0	1	1	0	0	0
Biology	6	2	4	2	4	1	1	0	0	0	2	0	2	0

Materials

Panelists evaluated the alignment of EOC intact test forms with the Missouri CLEs using the following materials.

Test Forms. Per EOC course, panelists evaluated test forms from each of the following administrations: Fall 2009, Summer 2010, and Spring 2011³. Table 2.2 lists the number and type of items per test form. Due to changes in test administration requirements, the 2010 and 2011 administrations only included multiple-choice (MC) items; thus, no performance events (PEs) were reviewed in the alignment study on the 2010 and 2011 test forms.

Table 2.2 Item Composition for Fall 2009, Summer 2010, and Spring 2011 EOC Test Forms

Course	Test Form														
	Fall 2009					Summer 2010					Spring 2011				
	Total Items	Operational		Field-test		Total Items	Operational		Field-test		Total Items	Operational		Field-test	
	MC	PE	MC	PE		MC	PE	MC	PE		MC	PE	MC	PE	
English II	48	35	1	12	0	47	35	0	12	0	47	35	0	12	0
Algebra I	49	35	2	12	0	47	35	0	12	0	47	35	0	12	0
Biology	62	35	15	12	0	47	35	0	12	0	47	35	0	12	0

Rating Forms and Instructions. Panelists rated the CLEs and test items by using Excel spreadsheet rating forms on laptops. Panelists reviewed the Missouri CLEs 2.0 for depth-of-knowledge (DOK) using a 4-point scale. For items, panelists made four judgments, including (a) item DOK, (b) content match to CLEs, (c) overall alignment rating on a 4-point scale, and (d) item quality rating on a 4-point scale. Panelists received instruction sheets listing the rating tasks and forms. Appendix E presents examples of rating forms and instructions.

Procedures

HumRRO directed the EOC alignment review November 2 and 3, 2010. The workshops began with introductions of staff and observers. Next, panelists read and signed affidavits of non-disclosure for the secure materials they would review during the workshop. HumRRO gave a presentation describing the purpose of the reviews and alignment research in general. This presentation briefly introduced the alignment tasks the panelists would perform.

Following the general introduction, panelists began working within their content groups. One HumRRO staff person experienced with alignment research facilitated the alignment process for each group. Within their small groups, facilitators further trained reviewers by instructing them on how to complete ratings and by answering questions on rating criteria. Group leaders provided general suggestions and comments when appropriate; however, they emphasized to reviewers that staff would not give explicit

³ This report does not include any examples of items or references to specific item content due to test security.

direction on how to rate standards or items because reviewers were valued as content experts. Each panelist worked at a computer station with access to Excel spreadsheets in which to enter ratings. Reviewers evaluated paper copies of EOC test items provided by the test vendor.

After completing training on DOK evaluations as a group, panelists proceeded to individually rate the CLEs relevant to each test. Once all reviewers completed their DOK ratings, groups discussed their ratings to achieve consensus on each CLE, which was recorded separately by the group leader.

Reviewers then received specific instructions on rating items. For training, group leaders led panelists in evaluating and discussing several sample items, followed by three to five items from an EOC assessment. Once the facilitator and group determined that reviewers were well-calibrated, reviewers began individual ratings of items starting with the Fall 2009 form, followed by the Summer 2010 and Spring 2011 forms.

Panelists assigned a primary CLE to an item based on their judgment that an item clearly measured this content. Reviewers could assign up to one additional CLE if they considered the item to assess another standard equally to the primary standard. Panelists completed all item ratings individually. After all reviewers completed a form, facilitators led panelists through an adjudication process on items with highly discrepant ratings. During the adjudication process, panelists were not required to reach consensus.

All reviewers finished tasks in approximately two days, although they completed their ratings at different times. After reviewers completed all three test forms, they provided summary comments about the degree of alignment. Finally, reviewers filled out a feedback survey on alignment training and process.

Statistical Method and Analysis

To reduce unnecessary repetition, this section presents a general description of the analyses performed as part of this alignment approach. While reviewers evaluated full intact test forms, we conducted all analyses on *operational* items only because these items compose student scores used in calculating Adequate Yearly Progress (AYP). The analyses include (a) inter-rater agreement, (b) four Webb alignment indicators, and (c) overall item ratings on alignment and quality. We present the results of these analyses per EOC course in the next three chapters (English II, Algebra I, and Biology).

For two of the three test forms, analyses cover selected-response operational items only (Summer 2010 and Spring 2011) because reviewers were asked to evaluate only multiple choice portions of the assessments. For the Fall 2009 form, reviewers did rate the operational performance events as well. We analyzed the Fall 2009 form with the performance events included, as well as with multiple choice items only. No field-test items were included in these alignment analyses.

Inter-rater Agreement Results

HumRRO performed two types of agreement analyses on reviewer alignment ratings. Reviewers rated the alignment of each item on two major dimensions: DOK and content match. The DOK rating required panelists to rank items using a scale, while the content rating involved a categorical judgment on the CLEs assessed by items. In each case, it is important to determine the extent to which panelists tended to provide exactly the same ratings on items (Shavelson, Webb, & Rowley, 1989; Tinsley & Weiss, 1975).

For item DOK ratings, Webb (2005) uses the intraclass correlation (ICC) coefficient. This type of agreement statistic involves the calculation of the ICC (C, k) statistic (Shrout & Fleiss, 1979). This statistic indicates the amount of agreement by producing a statistic between 0 and 1 (similar to a correlation coefficient). An ICC (C, k) result approaching 1 represents high agreement. Conversely, as the ICC approaches 0, we interpret this outcome to mean that panelists assigned quite different ratings to the same dimension, resulting in weak agreement. Generally, ICC outcomes can be interpreted based on the following decision criteria:

- Exact agreement 1.00
- Good agreement 0.80 to 0.99
- Adequate agreement 0.70 to 0.79
- Weak agreement 0.69 or less

Evaluating agreement between categorical ratings, such as CLEs matched to items, requires a different form of agreement statistic. Several agreement measures exist to analyze categorical ratings (see Gwet, 2001; Webb, 2005). Webb uses a statistic which basically estimates percent agreement between reviewers⁴. This analysis involves a pairwise comparison (one-to-one) of each reviewer's ratings with all other reviewers per item. Results are averaged across reviewers per test form. Webb's decision criteria for pairwise comparisons are comparable to those for the ICC, although calculations are slightly less stringent for exact agreement in particular.

- Exact agreement 1.00
- Good agreement 0.70 to 0.99
- Adequate agreement 0.60 to 0.69
- Weak agreement 0.59 or lower

Webb Alignment Measures

All of Webb's measures begin with calculations for each reviewer and progress to a summary of results across reviewers, usually per content *strand*. For each EOC test, we first calculated item frequency ratings per standard (CLE) for each panelist. Next, we calculated descriptive statistics (means and standard deviations) across reviewers. For

⁴ Refer to Webb, N. L. (2005). *Webb Alignment Tool (WAT): Training Manual* for a detailed discussion of the agreement analysis based on pairwise comparisons.

Algebra I and Biology, we generated mean ratings across panelists per strand. For English II, we computed means at the level of the Big Idea instead of Strand because the content is represented by a single strand level per domain (Reading and Writing), which can obscure some alignment issues.

Categorical Concurrence.

Categorical concurrence describes the extent to which EOC items cover the content strands in the Missouri CLEs. Webb recommends a minimum of six test questions to adequately assess each content strand. This criterion serves as a guideline for reasonable content coverage. This analysis involves determining the frequency of items matched to standards per strand per reviewers, then calculation of mean items per strand across reviewers.

DOK Consistency.

Analyses of DOK measure the type of cognitive processing required of students by content standards. These DOK requirements implied by the CLEs should be reflected in the corresponding assessment items. To confirm this match, the Webb method requires reviewers to separately rate the CLEs and the test items. Webb includes an alignment indicator, referred to as *depth-of-knowledge consistency*, that directly compares panelists' DOK ratings of content standards to their ratings of test items.

To make their ratings, panelists used the following rating scale (adapted from Webb, 2005) with four levels of cognitive complexity:

Table 2.3 Webb's Depth-of-knowledge Rating Scale

DOK Level/Title	DOK Description
Level 1 Recognition	Simple recall of information (i.e., facts, terms); sequencing; more automatic.
Level 2 Skills/Concepts	Beyond habitual response; applying concepts; problem-solving.
Level 3 Strategic Thinking	Requires basic reasoning, planning, or use of evidence; generating hypotheses.
Level 4 Extended Thinking	Complex reasoning; evaluation of multiple sources or independent pieces of evidence; often over an extended period of time.

HumRRO evaluated DOK consistency by analyzing the same ratings in two ways. One analysis focuses on the percent of items with appropriate DOK levels relative to corresponding CLEs, while the second analysis focuses on the percent of CLEs assessed at the correct DOK level by items. Thus, the DOK analyses are correlated, but the data presentation differs to highlight the assessment and the standards separately.

First, we determined the mean number of items with DOK below, at, and above the DOK level of the matched CLEs. These means were generated by calculating the frequency of items per reviewer at each DOK level relative to the corresponding standards. We summed item frequencies across CLEs per reviewer, and then calculated the mean number of items with DOK below, at, and above the standard DOK. We established the decision criterion that at least 50% of items per strand must match the DOK level of corresponding CLEs for acceptable alignment⁵.

Second, we calculated the mean number of CLEs per strand assessed below, at, and above the DOK level expected. For these calculations, we counted the number of CLEs where the assessed DOK for at least 50% of items fell below the standard DOK level, 50% of items assessed DOK at the same level as the CLE, and 50% assessed DOK above the corresponding CLEs.

Range-of-Knowledge.

The range-of-knowledge measure examines breadth of knowledge. In addition to evaluating which content strands are assessed, this measure considers how many of the CLEs within a strand are represented by items, with the guideline that the CLEs should be linked with at least one item. Webb's minimum level of acceptability for range-of-knowledge correspondence is that at least 50% of CLEs per strand link with items to ensure adequate breadth of content coverage *within* strands.

To determine how many of these CLEs were matched to items, we first computed the frequency of CLEs covered (per strand) separately for each panelist. Next, we calculated the mean number of CLEs linked with items across panelists.

Balance-of-Knowledge Representation.

The fourth measure of alignment included in the Webb method is *balance-of-knowledge representation*. This measure describes the distribution of items linked to each CLE within each strand. The number of items should be distributed relatively evenly between the CLEs to achieve good balance. However, the balance-of-knowledge results should be evaluated within the context of the state test blueprint, as well as the other three Webb alignment indicators.

The content balance is determined by calculating an index, or score, for each strand based on the number of items per CLE associated with that strand⁶. This index is based on item frequencies per CLE, which first are summed per reviewer. We then generated the mean frequency of items per CLE across reviewers for each strand. According to Webb, the minimum acceptable index for a single strand is 0.70 (on a

⁵ Webb's criterion is that DOK for 50% of items must be at *or above* corresponding content objective. HumRRO applies the criterion of requiring a match at the same level because assessing students above the level expected for proficiency also potentially assesses students inaccurately.

⁶ The exact formula for calculating the balance index is explained in detail in Webb's (2005) alignment training manual: <http://www.wcer.wisc.edu/WAT/index.aspx>.

scale of 0 to 1, with 1 representing perfect balance). An index of 0.70 or higher suggests that items broadly assess the CLEs matched to items by reviewers instead of clustering around one or two CLEs.

One point should be noted regarding the balance index when interpreting the results. Only those CLEs actually matched to items by the panelists are included in calculations of the balance index. A given strand may include more CLEs than are actually linked to items by panelists. For example, if a particular strand includes eight CLEs in the state content standards document but panelists found items matching to just three CLEs, only these three CLEs are evaluated for item distribution. Recognizing this feature of the balance index is important in cases when the range measure and balance measure produce seemingly contrasting results.

Chapter 3 Results: English II

In this chapter, we report results for English II including (a) inter-rater agreement and (b) summary results on the four Webb alignment indicators. At the end of this chapter, we highlight key outcomes. Detailed results can be found in Appendix A.

Inter-rater Agreement Results

In this section, we report on two types of agreement analyses concerning panelists' ratings. Refer to Chapter 2 for an explanation of these statistics and decision criteria. Table 3.1 presents inter-rater agreement outcomes (ICC) for item DOK ratings. These results are listed separately for the 2009, 2010, and 2011 test forms. The ICC (C, k) results in Table 3.1 indicate the reviewers consistently applied the same DOK ratings to the same items. All ICCs indicate 'Good agreement' between reviewers.

Table 3.1 Intraclass Correlation Coefficients on Item DOK Ratings for English II

Test Form	ICC Agreement Level
Fall 2009	0.98
Summer 2010	0.99
Spring 2011	0.99

Table 3.2 includes content match results at two levels of agreement. The first correlation presented for each form presents exact agreement results, reflecting agreement between reviewers at the Big Idea, Concept, and CLE level. The second correlation displayed for each test form indicates degree of partial agreement, reflecting an evaluation of agreement between reviewers at the Big Idea level only. Reviewers for English II were highly consistent in their determination of content assessed by items.

Table 3.2. Pairwise Comparisons on Reviewer Content Agreement for English II

Test Form	Exact Content Match (Big Idea, Concept, CLE)	Partial Content Match (Big Idea only)
Fall 2009	0.95	1.00
Summer 2010	0.92	1.00
Spring 2011	0.92	1.00

Webb Alignment Results

This section reviews the general outcomes of item analyses on the four Webb alignment indicators. As noted in Chapter 2, the Webb measures begin with an analysis of items per standard, then results are reported at the *strand* level. However, the content in Missouri's Communication Arts Grade Level Expectations and the Course

Level Expectations are divided into two broad content domains – Reading and Writing. As a result, it can be difficult to determine how well the breadth of content *within* these broad domains is assessed. Analysis with the Webb method can either over- or under-emphasize alignment issues with only two strands.

For this reason, HumRRO calculated analyses at the next level down within these strands, referred to in Missouri as “Big Ideas” for Communication Arts. Within Reading, for example, Missouri specifies three broad areas of content expectations: (a) Reading Processes, (b) Fiction, and (c) Nonfiction. Under Writing, Missouri expects students to know and demonstrate aspects of writing involved with: (a) the Writing Process, (b) Text Development, and (c) Forms/Types of Writing. We report the outcomes of the Webb alignment analyses for each Big Idea per strand.

We presented reviewers with the full set of Missouri CLEs for Reading and Writing. However, we note two qualifications to the assessment of the Writing strand. The Big Idea Forms/Types of Writing is intended for assessment mainly by student writing products (e.g., in-class reports; constructed response items). Thus, we would not expect to find many assessment items targeting this content. The Big Idea Writing-Processes’ primary intention is not standardized assessment; for this reason, we would not expect to find many items assessing this content. The outcomes of our analyses bear out these assumptions.

Under each Webb analysis, we present two sets of tables. In the first table, we present results on the 2009 test form first with multiple-choice and performance events (writing prompts for English II). The second table includes results on analyses of multiple-choice items only for the 2009, 2010, and 2011 test forms.

Categorical Concurrence

Tables 3.3 and 3.4 summarize the English II alignment results on categorical concurrence for test forms reviewed. Table 3.3, including results on the 2009 form with multiple-choice and performance events, indicates that all Big Ideas under the Reading strand met the minimum requirement of six items. While reviewers matched items to some Writing CLEs, the 2009 assessment does not cover a sufficient number of items to meet the alignment criterion per Big Idea. Reviewers determined that the assessment did not cover any CLEs under Writing-Process in particular, according to these reviewers. Note that the Writing strand was matched to a mean of 6.00 items overall.

Table 3.3 Summary of Categorical Concurrence Results for English II 2009 Test Form (Multiple-Choice and Performance Event Items)

Big Idea	Mean Items per Big Idea
Reading - Processes	12.40
Reading - Fiction	9.60
Reading - Nonfiction	7.80
^a Writing - Process	N/A
Writing - Text Development	5.00
^b Writing - Forms/Types	1.00
Big Ideas with at Least Six Items	
3 of 6	

Note: The *total* number of items matched to the Writing strand does meet the minimum requirement of six items.

Table 3.4 shows that all three test forms covered the Reading CLEs well. In contrast, the outcomes on Writing CLEs suggest that this content domain overall was not assessed well. The Missouri CLEs 2.0 intend for the Writing strand to be covered primarily by performance events on the assessment. Thus, the small number of multiple-choice items assessing writing is not unexpected given the structure of these content expectations (e.g., WR.1.A – “Apply a writing process to write effectively in various forms and types of writing”, Communication Arts Course Level Expectations 2.0, 2008).

Table 3.4 Summary of Categorical Concurrence Results for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Big Idea	Mean Items per Big Idea		
	2009 Test Form	2010 Test Form	2011 Test Form
Reading - Processes	12.60	13.40	11.60
Reading - Fiction	9.60	10.40	^c 5.60
Reading - Nonfiction	7.80	6.20	12.60
^a Writing - Process	0	0	0
Writing - Text Development	4.80	5.00	5.00
^b Writing - Forms/Types	0	0	0
Big Ideas with at Least Six Items			
3 of 6			

^a Reviewers did not match any items to Writing-Process.

^b No multiple-choice items matched to Writing-Forms/Types.

^c This outcome is marginally sufficient because it approaches the minimum criterion of six items per strand.

DOK Consistency

Tables 3.5 through 3.8 summarize the DOK consistency results for the English II test forms. Tables 3.5 and 3.6 focus on the mean percentage of *items* rated as matched to the CLEs on DOK level per Big Idea. Table 3.5 includes data for the 2009 test form with multiple-choice and performance event items. Over 50% of items assessed students at or above the level of cognitive complexity expected by the corresponding CLEs for Reading-Processes and Reading-Fiction. In contrast, only 37% of Reading-Nonfiction CLEs assessed students at the same cognitive level (no items assessed above CLEs). Thus, in most cases, inconsistency between the assessment and Reading CLEs on depth-of-knowledge occurred due to items at a lower level of complexity.

Of those items targeting Writing-Text Development, all assessed students at or above the level of the corresponding CLEs.

Table 3.5 Summary of DOK Consistency by Items for English II 2009 Test Form (Multiple-Choice and Performance Event Items)

Big Idea	Mean Percentage of Items At/Above DOK of CLEs
Reading - Processes	61%
Reading - Fiction	69%
Reading - Nonfiction	37%
Writing - Process	N/A
Writing - Text Development	100%
Writing - Forms/Types	0
Big Ideas Matched to 50% or More Items with Same DOK	
3 of 6	

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Table 3.6 includes results with multiple-choice items only for the 2009, 2010, and 2011 test forms. The same pattern emerged for all three test forms as noted in Table 3.5, although close to half of items on the 2010 and 2011 test forms did assess students at or above the DOK level of the corresponding CLEs. Again, of those Writing CLEs targeted, all items assessed students at a sufficient cognitive level.

Table 3.6 Summary of DOK Consistency by Items for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Big Idea	Mean Percentage of Items At/Above DOK of Corresponding CLEs		
	2009	2010	2011
Reading - Processes	61%	52%	63%
Reading - Fiction	69%	47%	40%
Reading - Nonfiction	37%	42%	41%
Writing - Process	N/A	N/A	N/A
Writing - Text Development	100%	100%	100%
Writing - Forms/Types	N/A	N/A	N/A
Big Ideas Matched to 50% or More Items with Same DOK	3 of 6	2 of 6	2 of 6

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Tables 3.7 and 3.8 present the same results according to the mean percentage of *CLEs* assessed appropriately (at the same DOK level) per Big Idea. When reviewing the results in this manner, the majority of targeted *CLEs* were assessed at a level comparable to what was expected. However, items assessing one *CLE* in particular (of three) under Reading-Nonfiction matched on DOK level more frequently. Specifically, items covering Informational and Persuasive Text assessed students at the same DOK level as expected by *over half of items*. In comparison, reviewers rated most items matched to the remaining two *CLEs* under Nonfiction as below standard on DOK.

Table 3.7 Summary of DOK Consistency by *CLEs* for English II 2009 Test Form (Multiple-Choice and Performance Event Items)

Big Idea	Number of <i>CLEs</i>	Mean Percentage of <i>CLEs</i> Assessed Appropriately by 50% of Items
Reading - Processes	3	72%
Reading - Fiction	3	73%
Reading - Nonfiction	3	35%
Writing - Process	1	N/A
Writing - Text Development	5	100%
Writing - Forms/Types	1	0
Big Ideas Assessed Appropriately		3 of 6

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Table 3.8 Summary of DOK Consistency by CLEs for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Big Idea	Number of CLEs	Mean Percentage of CLEs Assessed Appropriately by 50% of Items		
		2009	2010	2011
Reading - Processes	3	72%	70%	70%
Reading - Fiction	3	73%	52%	37%
Reading - Nonfiction	3	35%	42%	37%
Writing - Process	1	N/A	N/A	N/A
Writing - Text Development	5	100%	100%	100%
Writing - Forms/Types	1	N/A	N/A	N/A
Big Ideas Assessed Appropriately		3 of 6	3 of 6	2 of 6

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Range-of-Knowledge

Tables 3.9 and 3.10 summarize the range-of-knowledge results. For adequate coverage, one or more items should be matched to at least half of CLEs per strand⁷.

Table 3.9 includes results on the 2009 test form with multiple-choice and performance events. Panelists matched each Reading CLE (3 out of 3 per Big Idea) to at least one item, as demonstrated in the table by mean percentages of 100%. The 2009 form assessed only two Writing CLEs.

Table 3.9 Summary of Range-of-Knowledge Results for English II 2009 Test Form (Multiple-Choice and Performance Event Items)

Big Idea	Number of CLEs	Mean Percentage of CLEs Assessed by One or More Items
Reading - Processes	3	100
Reading - Fiction	3	100
Reading - Nonfiction	3	100
Writing - Process	1	0
Writing - Text Development	5	20
Writing - Forms/Types	1	^a 100
Big Ideas with Adequate Coverage		4 of 6

^a The single CLE was assessed by a single item.

⁷ This criterion refers to a unique item between test forms.

Table 3.10 presents results for all test forms with multiple-choice items only. The 2010 and 2011 test forms still met the minimum criterion for assessing the range of Reading CLEs with multiple-choice items only (at least two CLEs assessed). In contrast, reviewers matched only one Writing CLE to items; thus, elimination of the performance event decreased writing coverage.

Table 3.10 Summary of Range-of-Knowledge Results for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Big Idea	Number of CLEs	Mean Percentage of CLEs Assessed by One or More Items		
		2009	2010	2011
Reading - Processes	3	100	67	67
Reading - Fiction	3	100	80	60
Reading - Nonfiction	3	100	93	93
Writing - Process	1	0	0	0
Writing - Text Development	5	20	20	20
Writing - Forms/Types	1	0	0	0
Big Ideas with Adequate Coverage		3 of 6	3 of 6	3 of 6

Balance-of-Knowledge Representation

Tables 3.11 and 3.12 summarize the results on balance-of-knowledge representation for each test form. An index of 0.70 or higher indicates adequate distribution of items among assessed CLEs. As a reminder to the reader, the CLEs reflected in the balance indices only include those matched to items by panelists.

Of the assessed CLEs per Big Idea, the results in Table 3.11 suggest that items are distributed fairly evenly across this content. Note that, while the indices calculated Writing suggest a perfect distribution of items among CLEs, the Big Idea for Writing-Forms/Types includes only one CLE, which was assessed by one item (performance event). With the removal of the performance event, no balance index could be calculated for Writing-Forms/Types (see Table 3.12).

Table 3.11 Summary of Balance-of-Knowledge Results for English II 2009 Test Form (Multiple-Choice and Performance Event Items)

Big Idea	Balance Index per Strand
Reading - Processes	0.81
Reading - Fiction	0.83
Reading - Nonfiction	0.82
Writing - Process	N/A
Writing - Text Development	1.00
Writing - Forms/Types	1.00
Big Ideas Met Minimum Index	5 of 6

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Table 3.12 Summary of Balance-of-Knowledge Results for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Big Idea	Balance Index per Strand		
	2009	2010	2011
Reading - Processes	0.81	0.71	0.91
Reading - Fiction	0.83	0.75	0.91
Reading - Nonfiction	0.82	0.90	0.83
Writing - Process	N/A	N/A	N/A
Writing - Text Development	1.00	1.00	1.00
Writing - Forms/Types	N/A	N/A	N/A
Big Ideas Met Minimum Index	4 of 6	4 of 6	4 of 6

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Summary and Discussion of Webb Alignment Results for English II

The content alignment review evaluated the Fall 2009, Summer 2010, and Spring 2011 test forms for English II compared to the Missouri CLEs. Test forms aligned well to the Reading CLEs on content breadth. The assessments aligned partially to the Writing CLEs.

Summary alignment judgments displayed in Tables 3.13 and 3.14 are based on Webb (2005)⁸. These summary judgments focus on the percentage of content strands represented well by the assessment, and they reflect areas of strength and weakness (as opposed to a single, cumulative conclusion). Thus, these conclusions reflect a final

⁸ Tables 3.8 and 3.9 link to the bottom row of tables in Appendix A (Tables A-1 through A-10).

evaluation per Webb criteria across the Big Ideas per strand. Webb outlined a scale with a range of potential alignment outcomes applied to each of the four indicators:

- Fully aligned – assessments align to all content strands (100%)
- Highly aligned – assessments align to the majority of strands (70%–99%)
- Partially aligned – assessments align well to some strands (50%–69%)
- Weakly aligned – assessments align to less than half the strands (below 50%).

Table 3.13 includes the alignment conclusions for the 2009 test form based on analyses with multiple-choice and performance event items, while the conclusions in Table 3.14 are based on analyses of multiple-choice items only for all test forms.

Table 3.13 Summary Alignment Outcomes per Webb Criterion for English II 2009 Test Form with Multiple-Choice and Performance Event Items

English II Test Form	Alignment Conclusions per Webb Indicator			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Fall 2009	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (67%)	Highly aligned (80%)

Table 3.14 Summary Alignment Outcomes per Webb Criterion for English II 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

English II Test Forms	Alignment Conclusions per Webb Indicator			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Fall 2009	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)
Summer 2010	Partially aligned (50%)	Weakly aligned (37%)	Partially aligned (50%)	Partially aligned (50%)
Spring 2011	Partially aligned (50%)	Weakly aligned (37%)	Partially aligned (50%)	Partially aligned (50%)

These results require some explanation. First, the item DOK level appears to be inconsistent with the content expectations for some Reading items on each form, mostly relative to Nonfiction. Second, while some Reading items may warrant review, the overall alignment picture in Table 3.14 is impacted by the limited assessment of Writing in particular. As noted at the beginning of the chapter, this circumstance is to be expected because many Writing CLEs are intended for more comprehensive assessment through writing activities. If the analyses were conducted at the *strand* level, the outcomes for Reading would indicate a higher level of alignment, while the

outcomes for Writing would demonstrate lower alignment. Thus, all results must be taken within the context of the test blueprints for English II and design of the Missouri CLEs for Communication Arts.

Note that the 2009 test form exhibits stronger alignment to the CLEs on balance-of-representation when the performance event is included in the analyses. This outcome reflects the fact that the performance event did assess some additional writing content. However, all reviewers matched only two CLEs (of seven under the Writing strand) to items, so this result should be interpreted with caution.

Chapter 4 Results: Algebra I

In this chapter, we report results for Algebra I including (a) inter-rater agreement and (b) summary results on the four Webb alignment indicators. At the end of this chapter, we highlight key outcomes. Additional results can be found in Appendix B.

Inter-rater Agreement Results

In this section, we report on two types of agreement analyses concerning panelists' ratings. Refer to Chapter 2 for an explanation of these statistics and decision criteria. Table 4.1 presents inter-rater agreement outcomes (ICC) for item DOK ratings. These results are listed separately for the 2009, 2010, and 2011 test forms. The ICC (C, k) results in Table 4.1 indicate Algebra I reviewers rated item DOK in the same way in most cases. All ICCs indicate 'Good agreement' between reviewers.

Table 4.1 Intraclass Correlation Coefficients on Item DOK Ratings for Algebra I

Test Form	ICC Agreement Level
Fall 2009	0.94
Summer 2010	0.97
Spring 2011	0.92

Table 4.2 includes content match results at two levels of agreement. The first correlation presented for each form presents exact agreement results, reflecting agreement between reviewers across the board (Strand, Big Idea, Concept, and CLE level). The second correlation displayed for each test form indicates degree of partial agreement, reflecting an evaluation of agreement between reviewers at the Strand level only. Algebra I reviewers showed good consistency in items matched to CLEs. While some reviewers differed in exact match in several cases, all reviewers consistently agreed on the content strand assessed by items.

Table 4.2. Pairwise Comparisons on Reviewer Content Agreement for Algebra I

Test Form	Exact Content Match (Big Idea, Concept, CLE)	Partial Content Match (Big Idea only)
Fall 2009	0.88	1.00
Summer 2010	0.92	1.00
Spring 2011	0.92	1.00

Webb Alignment Results

This section reviews the general outcomes of item analyses on the four Webb alignment indicators. As noted in Chapter 2, the Webb measures begin with an analysis of items per standard, and then results are reported at the *strand* level.

Categorical Concurrence

Tables 4.3 and 4.4 summarize the Algebra I alignment results on categorical concurrence for the three test forms reviewed. Table 4.3 displays results on the 2009 form with multiple-choice and performance events. These results show that the test form represents the content of each of the strands. Reviewers matched twice as many items to the strand Algebraic Relationships, which corresponds with the content emphasis specified in the test blueprint. Furthermore, reviewers matched the performance events to Algebraic Relationships, as shown by the higher number of items ($M = 20.25$) compared to the analysis of the 2009 form with multiple-choice items only ($M = 18.25$) found in Table 4.4.

Table 4.3 Summary of Categorical Concurrence Results for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)

Strand	Mean Items per Strand
Numbers and Operations	9.25
Algebraic Relationships	20.25
Data and Probability	7.50
Strands with at Least Six Items	3 of 3

Table 4.4 includes results on all test forms based on an analysis of multiple-choice items only. Each test form covered the strands with a sufficient number of items to reflect overall breadth.

Table 4.4 Summary of Categorical Concurrence Results for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Mean Items per Strand		
	2009 Test Form	2010 Test Form	2011 Test Form
Numbers and Operations	9.25	8.75	9.50
Algebraic Relationships	18.25	18.00	19.25
Data and Probability	7.50	8.25	6.25
Strands with at Least Six Items	3 of 3	3 of 3	3 of 3

DOK Consistency

Tables 4.5 and 4.8 summarize the DOK consistency results for the Algebra I test forms. Tables 4.5 and 4.6 display the mean percentage of items matched to corresponding CLEs on DOK level. Table 4.5 (2009 test form with all operational items) reveals that well over half of items assess students at the same or higher level of complexity as the targeted CLE.

Table 4.5 Summary of DOK Consistency by Items per Test Form for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)

Strand	Mean Percentage of Items with DOK At/Above CLE
Numbers and Operations	85%
Algebraic Relationships	78%
Data and Probability	58%
Strands Matched to 50% or More Items with Same DOK	3 of 3

Table 4.6 shows that the majority of items on each test form assess students at the same DOK level as the corresponding CLEs relative to the Numbers and Operations and the Algebraic Relationships strands. However, considerably fewer items matched the DOK level of the Data and Probability CLEs for the 2010 and 2011 test forms in particular. The depth-of-knowledge expected by the CLEs under Data Probability should be between DOK level 2 to 3, while a number of items were rated as DOK level 1 or 2 by these reviewers. The discrepancy between the DOK of items and CLEs can be categorized as ‘adjacent’ (item DOK=2 vs. CLE DOK=3) for nearly all items, which is less concerning than if ratings differed by two levels (i.e., item DOK=1 vs. CLE DOK=3). Furthermore, it is reasonable that at least a few items included under this strand would assess student knowledge at DOK level 1 to better discriminate student performance.

Table 4.6 Summary of DOK Consistency by Items per Test Form for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Mean Percentage of Items with DOK At/Above CLE		
	2009	2010	2011
Numbers and Operations	85%	94%	71%
Algebraic Relationships	75%	88%	74%
Data and Probability	58%	39%	23%
Strands Matched to 50% or More Items with Same DOK	3 of 3	2 of 3	2 of 3

Table 4.7 presents the data according to the mean percent of CLEs assessed at the cognitive level expected based on reviewer DOK consensus values. These results, of course, follow the same pattern as noted in Table 4.5 focusing on items.

Table 4.7 Summary of DOK Consistency by CLEs per Test Form for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)

Strand	Number of CLEs	Mean Percentage of CLEs Assessed At/Above DOK Expected by 50% of Items
Numbers and Operations	2	88%
Algebraic Relationships	10	79%
Data and Probability	5	71%
Strands Assessed Appropriately		3 of 3

As shown in Table 4.8, most CLEs under the Numbers and Operations and Algebraic Relationships strands received appropriate assessment by at least half of matched items. Corresponding with Table 4.6, only one or two Data and Probability CLEs were assessed at the appropriate DOK level.

Table 4.8 Summary of DOK Consistency by CLEs per Test Form for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Number of CLEs	Mean Percentage of CLEs Assessed At/Above DOK Expected by 50% of Items		
		2009	2010	2011
Numbers and Operations	2	88%	100%	88%
Algebraic Relationships	10	78%	84%	71%
Data and Probability	5	71%	44%	25%
Strands Assessed Appropriately		3 of 3	2 of 3	2 of 3

Range-of-Knowledge

Tables 4.9 and 4.10 summarize the range-of-knowledge results for each test form. At least 50% of CLEs per strand should be assessed by one or more items for adequate coverage. Table 4.9 shows that panelists matched operational items on the 2009 form to almost all CLEs. Similarly, the multiple-choice portion of the three test forms assessed the range of the CLEs well, as show in Table 4.10.

Table 4.9 Summary of Range-of-Knowledge Results for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)

Strand	Number of CLEs	Mean Percentage of CLEs Assessed by One or More Items
Numbers and Operations	2	100%
Algebraic Relationships	10	90%
Data and Probability	5	85%
Strands with Adequate Coverage		3 of 3

Table 4.10 Summary of Range-of-Knowledge Results for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Number of CLEs	Mean Percentage of CLEs Assessed by One or More Items		
		2009	2010	2011
Numbers and Operations	2	100%	100%	100%
Algebraic Relationships	10	90%	95%	78%
Data and Probability	5	85%	90%	85%
Strands with Adequate Coverage		3 of 3	3 of 3	3 of 3

Balance-of-Knowledge Representation

Tables 4.11 and 4.12 show the results of balance-of-knowledge representation calculations for each test form. An index of 0.70 or higher indicates adequate distribution of items among assessed CLEs. The items associated with each strand met this minimum for each test form, although the resulting indices for the 2010 and 2011 test forms were slightly lower compared to the 2009 form.

Table 4.11 Summary of Balance-of-Knowledge Results for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items Only)

Strand	Balance Index per Strand	
Numbers and Operations	0.98	
Algebraic Relationships	0.96	
Data and Probability	0.81	
Balance Index Met per Strand		3 of 3

Table 4.12 Summary of Balance-of-Knowledge Results for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Balance Index per Strand		
	2009	2010	2011
Numbers and Operations	0.98	0.74	0.82
Algebraic Relationships	0.98	0.75	0.74
Data and Probability	0.81	0.81	0.83
Balance Index Met per Strand	3 of 3	3 of 3	3 of 3

Summary and Discussion of Webb Alignment Results for Algebra I

The content alignment review evaluated the Fall 2009, Summer 2010, and Spring 2011 test forms for Algebra I compared to the Missouri CLEs. Overall, the results suggest that these Algebra I test forms aligned well to the CLEs.

Summary alignment judgments displayed in Tables 4.8 and 4.9 are based on Webb (2005)⁹. These summary judgments focus on the percentage of content strands represented well by the assessment, and they reflect areas of strength and weakness (as opposed to a single, cumulative conclusion). Thus, these conclusions reflect a final evaluation per Webb criteria across strands. Webb outlined a scale with a range of potential alignment outcomes applied to each of the four indicators:

- Fully aligned – assessments align to all content strands (100%)
- Highly aligned – assessments align to the majority of strands (70%–99%)
- Partially aligned – assessments align well to some strands (50%–69%)
- Weakly aligned – assessments align to less than half the strands (below 50%).

Table 4.13 includes the alignment conclusions for the 2009 test form based on analyses with multiple-choice and performance event items, while the conclusions in Table 4.14 are based on analyses of multiple-choice items only for all test forms. Clearly, the outcomes of the analyses for the 2009 test form with and without the performance events are the same.

⁹ Tables 4.8 and 4.9 link to the bottom row of tables in Appendix B (Tables B-1 through B-10).

Table 4.13 Summary Alignment Outcomes per Webb Criterion for Algebra I 2009 Test Form (Multiple-Choice and Performance Event Items)

Algebra I Test Form	Alignment Conclusions per Webb Indicator			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Fall 2009	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)

Table 4.14 Summary Alignment Outcomes per Webb Criterion for Algebra I 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Algebra I Test Forms	Alignment Conclusions per Webb Indicator			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Fall 2009	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)
Summer 2010	Fully aligned (100%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)
Spring 2011	Fully aligned (100%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)

As evident in Table 4.14, some items included on the 2010 and 2011 assessments appear to assess students at a lower depth-of-knowledge level than expected. This conclusion pertains to some items designed to assess the Data and Probability CLEs. While the cognitive complexity of these items may warrant review, the degree of discrepancy between items and CLEs is off by one DOK level in almost all instances. Furthermore, at least three of the CLEs under Data and Probability expect students to demonstrate knowledge at DOK level 3, a high level of processing. Assessments generally include at least several items at lower levels of cognitive complexity per strand to further discriminate among achievement.

Recommendations and suggestions for improving alignment between the Algebra I assessments and Missouri CLEs are discussed in Chapter 6 (Summary and Recommendations).

Chapter 5 Results: Biology

In this chapter, we report results for Biology including (a) inter-rater agreement and (b) summary results on the four Webb alignment indicators. At the end of this chapter, we highlight key outcomes.

As a preface to the results in this section, we point out that the Scientific Inquiry strand is unique for two reasons. First, it is a process strand intended for assessment along with content strands. Second, DESE intends for Scientific Inquiry to be assessed at the State-level by performance events, which is evident in the test blueprints. As a result, the assessments cover Scientific Inquiry in a more limited way overall, compared to content strands. Furthermore, the State suspended assessment of student knowledge by performance events starting in 2010. Thus, the 2010 and 2011 test forms do not explicitly intend to assess students on this knowledge. We included the Scientific Inquiry strand in data analysis; however, we provide explanation and reminders to the reader relative to this strand for the 2010 and 2011 results in particular.

Additional results can be found in Appendix B.

Inter-rater Agreement Results

In this section, we report on two types of agreement analyses concerning panelists' ratings. Refer to Chapter 2 for an explanation of these statistics and decision criteria. Table 5.1 presents inter-rater agreement outcomes (ICC) for item DOK ratings. These results are listed separately for the 2009, 2010, and 2011 test forms. The ICC (C, k) results in Table 5.1 indicate that most Biology reviewers rated item DOK in the same way. One reviewer deviated from the group more often than other reviewers. However, all ICCs still indicate 'Good agreement' between reviewers.

Table 5.1 Intraclass Correlation Coefficients on Item DOK Ratings for Biology

Test Form	ICC Agreement Level
Fall 2009	0.88
Summer 2010	0.91
Spring 2011	0.82

Table 5.2 includes content match results at two levels of agreement. The first correlation presented for each form presents exact agreement results, reflecting agreement between reviewers across the board (Strand, Big Idea, Concept, and CLE level). The second correlation displayed for each test form indicates degree of partial agreement, reflecting an evaluation of agreement between reviewers at the Strand level only. Biology reviewers showed good consistency in items matched to CLEs. While some reviewers differed in exact match in several cases, all reviewers consistently agreed on the content strand assessed by items. Reviewers seemed to have the most

difficulty in agreeing on CLEs matched to items for the Spring 2011 test form. A result of 0.78 reflects adequate agreement between reviewers.

Table 5.2. Pairwise Comparisons on Reviewer Content Agreement for Biology

Test Form	Exact Content Match (Big Idea, Concept, CLE)	Partial Content Match (Big Idea only)
Fall 2009	0.82	1.00
Summer 2010	0.91	1.00
Spring 2011	0.78	1.00

Webb Alignment Results

This section reviews the general outcomes of item analyses on the four Webb alignment indicators. As noted in Chapter 2, the Webb measures begin with an analysis of items per standard; results are reported at the *strand* level.

Categorical Concurrence

Tables 5.3 and 5.4 summarize the Biology alignment results on categorical concurrence for test forms reviewed. Table 5.3 includes results of analyses on the 2009 form with multiple-choice and performance events. When all 2009 operational items were included, all three Biology strands received adequate representation.

Table 5.3 Summary of Categorical Concurrence Results for Biology 2009 Test Forms (Multiple-Choice and Performance Event Items)

Strand	Mean Items per Big Idea
Living Organisms	21.83
Ecology	13.00
Scientific Inquiry	14.83
Strands with At Least Six Items	3 of 3

In comparison, analyses on the multiple-choice portion of the three test forms, displayed in Table 5.4, revealed that reviewers did not find any items clearly matched the Scientific Inquiry strand. This outcome does correspond with statements in the test specifications document regarding the assessment of Scientific Inquiry with performance events.

Table 5.4 Summary of Categorical Concurrence Results for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Mean Items per Big Idea		
	2009 Test Form	2010 Test Form	2011 Test Form
Living Organisms	21.83	21.64	22.00
Ecology	13.00	13.46	13.00
Scientific Inquiry	0	0	0
Strands with At Least Six Items	2 of 3	2 of 3	2 of 3

DOK Consistency

Tables 5.5 and 5.8 summarize the DOK consistency results for the Biology test forms. Tables 5.5 and 5.6 focus on the proportion of items matched to the CLEs on DOK.

As shown in Table 5.5, well over half of items assessing CLEs under Living Organisms and Ecology matched on DOK level. In contrast, few items ($M = 2.1$) assessed Scientific Inquiry CLEs at the appropriate DOK level.

Table 5.5 Summary of DOK Consistency by Items per Test Form for Biology 2009 Test Forms (Multiple-Choice and Performance Event Items)

Strand	Mean Percentage of Items with Same DOK as Corresponding CLEs
Living Organisms	77
Ecology	69
Scientific Inquiry	14
Strands Matched to 50% or More Items with Same DOK	2 of 3

Analyses on the three test forms with multiple-choice operational items produced similar results (Table 5.6).

Table 5.6 Summary of DOK Consistency by Items per Test Form for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Mean Percentage of Items with Same DOK as Corresponding CLEs		
	2009	2010	2011
Living Organisms	77	81	69
Ecology	69	81	78
Scientific Inquiry	N/A	N/A	N/A
Strands Matched to 50% or More Items with Same DOK	2 of 3	2 of 3	2 of 3

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Tables 5.7 and 5.8 present the same data according to the number of CLEs assessed appropriately. The DOK levels of over half of CLEs were assessed appropriately for Living Organisms and Ecology strands. Only two CLEs were assessed at the expected cognitive level for Scientific Inquiry. Six CLEs (of eight assessed) exhibit DOK Level 3. In contrast, reviewers rated most performance events ($M = 10.32$ items) as assessing students at DOK Level 2 and remaining items at DOK Level 1.

Table 5.7 Summary of DOK Consistency by CLEs per Test Form for Biology 2009 Test Forms (Multiple-Choice and Performance Event Items)

Strand	Number of CLEs	Mean Percentage of CLEs Assessed Appropriately by 50% of Items
Living Organisms	17	83%
Ecology	8	72%
Scientific Inquiry	15	14%
Strands Assessed Appropriately		2 of 3

As shown in Table 5.8, the multiple-choice portion of the assessment matched the DOK level of the CLEs for the majority of items. Thus, these test forms required students in Biology I to demonstrate knowledge at the same level as expected in the CLEs for Living Organisms and for Ecology.

Table 5.8 Summary of DOK Consistency by CLEs per Test Form for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strand	Number of CLEs	Mean Percentage of CLEs Assessed Appropriately by 50% of Items		
		2009	2010	2011
Living Organisms	17	83	75	70
Ecology	8	72	81	80
Scientific Inquiry	15	N/A	N/A	N/A
Strands Assessed Appropriately		2 of 3	2 of 3	2 of 3

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Range-of-Knowledge

Tables 5.9 and 5.10 summarize the range-of-knowledge results for the Biology test forms reviewed. At least 50% of CLEs per strand should be assessed by one or more items for adequate coverage. The Living Organisms and Ecology strands met this requirement.

Table 5.9 Summary of Range-of-Knowledge Results for Biology 2009 Test Form (Multiple-Choice and Performance Event Items)

Strands	Number of CLEs	Mean Percentage of CLEs Assessed by One or More Items
Living Organisms	17	82%
Ecology	8	88%
Scientific Inquiry	15	39%
Strands with Adequate Coverage		2 of 3

Table 5.10 Summary of Range-of-Knowledge Results for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strands	Number of CLEs	Mean Percentage of CLEs Assessed by One or More Items		
		2009	2010	2011
Living Organisms	17	82%	70%	80%
Ecology	8	88%	98%	73%
Scientific Inquiry	15	0	0	0
Strands with Adequate Coverage		2 of 3	2 of 3	2 of 3

Balance-of-Knowledge Representation

Tables 5.11 and 5.12 display results on balance-of-knowledge representation. An index of 0.70 or higher indicates adequate distribution of items among assessed CLEs. For the 2009 test form with all operational items, the balance indices reflect good item distribution among CLEs for each content strand. Recall, however, from Table 5.9 that reviewers matched performance events to only 39% ($M = 5.85$) of CLEs under Scientific Inquiry; thus, the balance index reflects the distribution of performance events among these assessed CLEs.

Table 5.11 Summary of Balance-of-Knowledge Results for Biology 2009 Test Form (Multiple-Choice and Performance Event Items)

Strands	Balance Index per Strand
Living Organisms	0.80
Ecology	0.84
Scientific Inquiry	0.76
Balance Index Met per Strand	3 of 3

As shown in Table 5.12, multiple-choice items on the 2009, 2010, and 2011 forms demonstrated good balance of content relative to the CLEs for Living Organisms and for Ecology.

Table 5.12 Summary of Balance-of-Knowledge Results for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Strands	Balance Index per Strand		
	2009	2010	2011
Living Organisms	0.80	0.77	0.82
Ecology	0.84	0.91	0.88
Scientific Inquiry	N/A	N/A	N/A
Balance Index Met per Strand	2 of 3	2 of 3	2 of 3

Note: N/A indicates that analysis was not conducted because no items were matched to this content.

Summary and Discussion of Webb Alignment Results for Biology

The content alignment review evaluated the Fall 2009, Summer 2010, and Spring 2011 test forms for Biology compared to the Missouri CLEs. The test forms aligned well to the Living Organism and Ecology strands on breadth and depth. The assessments do not cover the Scientific Inquiry strand as well, an outcome that is partly expected and

explainable. Analyses on the 2009 form with performance events show that these items covered a small portion of the total CLEs for Scientific Inquiry available for state-level assessment. However, the absence of assessment items covering Scientific Inquiry on the 2010 and 2011 test forms can be attributed entirely to the elimination of performance events, which corresponds with the test specifications document.

Summary alignment judgments displayed in Tables 5.8 and 5.9 are based on Webb (2005)¹⁰. These summary judgments focus on the percentage of content strands represented well by the assessment, and they reflect areas of strength and weakness (as opposed to a single, cumulative conclusion). Thus, these conclusions reflect a final evaluation per Webb criteria *across* the strands. Webb outlined a scale with a range of potential alignment outcomes applied to each of the four indicators:

- Fully aligned – assessments align to all content strands (100%)
- Highly aligned – assessments align to the majority of strands (70%–99%)
- Partially aligned – assessments align well to some strands (50%–69%)
- Weakly aligned – assessments align to less than half the strands (below 50%).

Table 5.13 includes the alignment conclusions for the 2009 test form based on analyses with multiple-choice and performance event items, while the conclusions in Table 5.14 are based on analyses of multiple-choice items only for all test forms.

Table 5.13 Summary Alignment Outcomes per Webb Criterion for Biology 2009 Test Form (Multiple-Choice and Performance Event Items)

Biology Test Form	Alignment Conclusions per Webb Indicator			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Fall 2009	Fully aligned (100%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)

¹⁰ Tables 5.8 and 5.9 link to the bottom row of tables in Appendix C (Tables C-1 through C-10).

Table 5.14 Summary Alignment Outcomes per Webb Criterion for Biology 2009, 2010, and 2011 Test Forms (Multiple-Choice Items Only)

Test Forms	Alignment Conclusions per Webb Indicator			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Fall 2009	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)
Summer 2010	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)
Spring 2011	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)

Relative to the Living Organisms and Ecology strands, each test form includes well over the minimum number of items needed (solely based on selected-response items) to represent the breadth of this content. Furthermore, items target a broad range of CLEs for these two strands. Finally, the majority of multiple-choice items assessed student knowledge at the level of complexity designated by the corresponding CLEs, although some performance events assessed students at a lower level than expected (i.e., Item DOK Level = 1 or 2 and CLE DOK = 3).

The lower degree of alignment across Webb measures indicated in Table 5.14 occurred primarily due to the absence of performance events in these alignment analyses. The test specifications document for Biology indicates that test forms assess the Scientific Inquiry strand through performance events, although the CLEs 2.0 document does not make this same statement. Thus, without the inclusion of performance events, the 2010 and 2011 test forms do not assess this strand.

Suggestions for improving the alignment between the Biology assessments and Missouri CLEs are discussed in Chapter 6 (Summary and Recommendations).

Chapter 6: Summary and Recommendations

HumRRO conducted forms alignment validation studies of the Missouri Assessment Program End-of-Course test forms for English II, Algebra I, and Biology relative to the Missouri CLEs. Reviewers evaluated a single test form from each of three administrations: Fall 2009, Summer 2010, and Spring 2011. Alignment of assessments and achievement standards to state content standards is a requirement of NCLB legislation.

The extent of alignment to the Missouri CLEs varied per content area and test form. In terms of confirmatory alignment evidence, all of the English II test forms covered the breadth of the Reading CLEs quite well. The Algebra I results suggest that the test forms fully align to the breadth of the CLEs across strands. For Biology, all test forms exhibited full alignment on content breadth and depth relative to the Living Organisms and Ecology strands.

The test forms reviewed for each course also exhibited some gaps that DESE may wish to review to improve alignment. Regarding English II, we point to two potential alignment issues. First, the depth-of-knowledge assessed does not match the CLEs for many items targeting Reading-Nonfiction. Second, the test forms assess Writing content in a narrow way, an outcome that can be explained in part by the design of the test blueprint and State suspension of performance events (Session II). Assessing writing skills on a state-level assessment can be challenging, particularly if no writing component is in place as is currently the case. As a result, reviewers determined that the selected-response portion of the assessment covers a single Writing CLE (Text Development - Conventions of English) with approximately five items. With the performance event (included in the Fall 2009 administration), reviewers still matched items to only two CLEs total (Conventions of English and Forms/Types of Writing). Thus, the test forms reviewed as part of this study partly align to the content standards.

For Algebra I, some test items on the 2010 and 2011 test forms may require review of assessed DOK for the Data and Probability strand. Reviewers rated over half of items on these forms as below the DOK level of the targeted CLEs. In most cases, the degree of discrepancy involved an adjacent mismatch (i.e., item DOK=2; CLE DOK=3). The 2009 test form did surpass the minimum criterion (M=58% of items matched the DOK of the CLEs).

The Biology test forms show gaps specifically in assessment of Scientific Inquiry in breadth and depth. The limited assessment of the breadth of this strand overall does correspond with the design of Missouri CLEs for Biology, however. The Scientific Inquiry strand is intended for assessment only by performance events because it requires students to demonstrate integrated understanding of scientific principles, particularly the application of experimental procedures. Given that the State was forced to eliminate performance events from the 2010 and 2011 Biology test forms, this circumstance makes it challenging for DESE to assess this content at this time. Furthermore, we recognize that these content expectations serve as a process strand to be assessed

along with other science content strands. DESE will want to determine how to handle this strand as they transition to a new test vendor.

We do note that, for the 2009 test form with performance events, the depth of content assessed did not match the CLEs under Scientific Inquiry for some performance events. Analyses of the 2009 form with performance events show that the majority of performance items assess students at a lower level of cognitive complexity than expected (i.e., item DOK=2 and CLE DOK=3). Thus, if DESE can pursue state-level assessment of Scientific Inquiry in the future, an increase in cognitive complexity may be needed.

Tables 6.1 and 6.2 provide summary alignment conclusions for each course assessment per Webb alignment indicator. The conclusions in Table 6.1 focus on the alignment analyses on the 2009 test forms for each course, including multiple-choice and performance event items. In comparison, Table 6.2 only displays results for analyses on the multiple-choice items.

Table 6.1. Summary Alignment Conclusions for 2009 Test Form per Course by Webb Alignment Indicator (Multiple-Choice and Performance Event Items)

Test Forms	Alignment Conclusions per Webb Indicator			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
English II	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (67%)	Highly aligned (80%)
Algebra I	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)
Biology	Fully aligned (100%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)

Table 6.2. Summary Alignment Conclusions per Course Test Form by Webb Alignment Indicator (Multiple-Choice Items Only)

Test Forms	Alignment Conclusions per Webb Indicator											
	Categorical Concurrence			Depth-of-Knowledge Consistency			Range-of-Knowledge Correspondence			Balance-of-Knowledge Representation		
	Fall 2009	Summer 2010	Spring 2011	Fall 2009	Summer 2010	Spring 2011	Fall 2009	Summer 2010	Spring 2011	Fall 2009	Summer 2010	Spring 2011
English II	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Weakly aligned (37%)	Weakly aligned (37%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)	Partially aligned (50%)
Algebra I	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Partially aligned (67%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)
Biology	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)	Partially aligned (67%)

Based on these results, HumRRO offers several recommendations to DESE on ways in which test alignment might be improved. We recognize that even minor changes to operational items require time for implementation. Thus, we would expect any modifications to items or standards to occur over the course of a normal review cycle (two to three years).

We also suggest that DESE, along with the test developer, review the results and recommendations relative to the test blueprints to determine if some outcomes per EOC content test are justifiable.

Recommendations

English II

1. **Review DOK assessed by Reading items relative to the Missouri CLEs (DOK consistency).** Panelists' ratings of item DOK for each test form indicate that some items assess student knowledge below the DOK level of corresponding CLEs. This was particularly true for 2010 and 2011 items targeting Nonfiction. Changing DOK for as few as three items covering Reading would improve alignment above the minimum criterion.
2. **Review the content distribution of items assessing Writing CLEs.** While the test blueprint does specify that many Writing CLEs should be assessed by performance events, DESE may wish to determine if writing could be assessed in a more even manner. If suspension of performance events continues to be a necessity, one approach in future administrations for increasing alignment could be to include additional selected-response format items requiring text evaluation for writing structure. We recognize that this solution may involve item development, which may be cost prohibitive in the immediate future. A second option may be to tie course-level student writing products to the state assessment. A few states have pursued this option by including writing components graded by teachers, based on a state-developed rubric, as part of scores. Finally, DESE should (and probably has already) emphasize to teachers that course-level assessment of writing is critical to ensure sufficient coverage of these skills.

Algebra I

1. **Review DOK for some items on 2010 and 2011 test forms relative to the Data and Probability strand (DOK consistency).** Reviewers' ratings indicate a lower level cognitive complexity for items assessing the Data and Probability strand, particularly for the 2010 and 2011 test forms. As noted for English II, improving DOK alignment could involve minor item edits to stems and/or response options for one to three items. However, we recognize that discrepancy between items and CLEs for this strand is inevitable given that these particular CLEs, and Data and Probability content in general, require

students to demonstrate knowledge at a higher level of processing.

Biology

1. **Review the test forms for coverage of the Scientific Inquiry strand (all Webb indicators).** While the test forms align very well to the Living Organisms and Ecology strands, reviewers found substantial gaps in assessment of Scientific Inquiry. However, this issue is expected, as noted in the earlier discussion. First, this strand is intended for assessment only by performance events because it requires students to demonstrate integrated understanding of scientific principles, particularly the application of experimental procedures. Thus, the limited assessment of Scientific Inquiry, even on the 2009 test form including performance events, does accurately reflect the intention of the standards and the test blueprint. Second, and consequently, the elimination of performance events from Biology test forms makes it challenging for the state to assess this content at this time.

If circumstances change for DESE and additional funds become available, we offer several recommendations for item development. In a similar way as recommended for the Writing strand for English II, we suggest considering whether some CLEs under this strand could be assessed by selected-response items. As an embedded strand, selected-response items could address scientific inquiry along with a primary content strand, which may be possible with current items. Alternatively, DESE could pursue item development of basic knowledge of the scientific process. While not ideal given that the intention of this strand is to encourage student reasoning and analysis, some representation of this strand would achieve greater breadth of the State-level content expectations.

References

- Brennan, R. L. (2001). *Generalizability theory* (2nd ed.). New York: Springer.
- Brennan, R. L. & Kane, M.T. (1977). Signal/noise ratios for domain-referenced tests. *Psychometrika*, 42(4), 609-625
- Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD: STATAXIS.
- No Child Left Behind Act of 2001. Public Law 107-110.
- Putka, D. & Sackett, P. (in press). Reliability and validity. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of Employee Selection* (pp. 9-49). London: Psychology Press.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Shavelson, R. J., Webb, N. M., & Rowley, G.L. (1989). Generalizability theory. *American Psychologist*, 44(6), 922-932.
- Tinsley, H. E. & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358-376.
- U. S. Department of Education. (April, 2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education. Retrieved from <http://www.ed.gov/policy/elsec/guid/saaprguidance.doc>.
- Webb, N. L. (2005). *Webb alignment tool: Training manual*. Madison, WI: Wisconsin Center for Education Research. Available: <http://www.wcer.wisc.edu/WAT/index.aspx>.
- Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in Algebra I and Biology education*. Washington, D.C.: Council of Chief State Schools Officers.
- Webb, N. L. (1999). *Alignment of Biology and Algebra I standards and assessments in four states. (Research Monograph 18)*. Madison, WI: National Institute for Biology Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852).

