

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

Leslie R. Taylor
Norman L. Webb, subcontractor
Milton E. Koger
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: C308004001-002

December 30, 2009

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

**Leslie R. Taylor
Norman L. Webb, subcontractor
Milton E. Koger
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: C308004001-002

December 30, 2009

EXECUTIVE SUMMARY

Scope of Work

The Missouri Department of Elementary and Secondary Education (DESE) requested an external independent alignment study of the End-of-Course (EOC) assessments for English II, Algebra I, and Biology. Specifically, the study evaluated the alignment of the Spring 2009 (Form 1) and Summer 2009 (Form 2) EOC test forms to the Missouri Course-Level Expectations (CLEs)¹. Missouri uses the EOC tests in the federal and state accountability programs. DESE awarded the Human Resources Research Organization (HumRRO) the contract to conduct this alignment study, along with Dr. Norman Webb as subcontractor.

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 panels of Missouri educators and national content experts to review the EOC test forms. These panelists included current and former teachers, administrators, and curriculum specialists or district coordinators. Each panel included both in-state and out-of-state panelists.

Dr. Norman Webb directed this alignment review at the Assessment Resource Center (ARC) at the University of Missouri, Columbia, on July 15 and 16, 2009. While panels were convened in facilities procured through DESE, Dr. Webb directed the actual reviews independently of DESE. Dr. Webb used the alignment method he developed (1997; 1999; 2005) to evaluate the alignment of the Spring 2009 and Summer 2009 EOC test forms for English II, Algebra I, and Biology to the Missouri Course-Level Expectations. As part of this method, reviewers rate individual test items

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

on the cognitive complexity and content assessed relative to the Missouri Course- Level Expectations. Dr. Webb’s 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) 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

Key Findings and Conclusions

For English I, a number of outcomes point to strong content alignment of the EOC to the Missouri Course-Level Expectations (CLEs). Each form reviewed clearly includes a sufficient number of operational test items to cover the major content categories (strands), as demonstrated by the outcomes on categorical concurrence. Panelists found items matching a sufficient number of CLEs per strand, indicating that the assessment covers reasonable breadth of content. Furthermore, the balance-of-knowledge representation results suggest that items seem to be distributed reasonably, at least across CLEs matched by panelists. However, two features of the test forms may warrant review. First, the DOK level of items assessing Reading (for Form 1 in particular) should be increased for approximately half of the items to better match the CLEs, as noted by the conclusion of ‘partially aligned.’ Second, while the range-of-knowledge correspondence outcomes produced a final judgment of ‘fully aligned’ based on the Webb minimum criterion, both test forms assessed a relatively narrow range of CLEs (impact is greater for Writing). This issue, in conjunction with the finding regarding item distribution, suggests that DESE may wish to review content emphasis on the assessment. A disproportionate emphasis of some content may be intended by DESE, which could be confirmed and justified by the test blueprint. For those CLEs matched to only one item, however, DESE may consider whether this number is sufficient to demonstrate accurate assessment of student knowledge of these content expectations.

The overall alignment results for the Algebra I test forms suggest that test items align well to the CLEs on breadth of content coverage. However, items on Form 2 (Summer) do not meet the DOK requirements of the CLEs. Specifically, panelists found that less than 50% of items assessed students at the same cognitive levels expected for

the CLEs under the Numbers and Operations and Algebraic Relationships strands. It also should be noted that exactly 50% of items met depth requirements for the Data and Probability strand. This third CLE was at the minimum percentage required.

The overall alignment results for the Biology test forms were mixed. The 2009 test forms include a sufficient number of items to adequately cover the breadth of the Science content strands for Biology. However, items target a narrow range of CLEs for Scientific Inquiry in particular, with only 42% of the CLEs within the strand covered by an item. While there were 20 items matched to this strand's CLEs, panelists matched only six of this strand's 15 CLEs with an item. As a result of the findings for range-of-knowledge, the findings on balance-of-knowledge representation should be interpreted with caution. In addition, many items (over half for two of three strands) assess student knowledge of the content at a lower level of cognitive depth than required by corresponding CLEs. This results in a finding of 'weakly aligned' for this measure.

Alignment of EOC Test Forms to Missouri Course-Level Expectations

Table 1 provides summary conclusions on the alignment of the EOC 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%).

Table 1. Summary Alignment Conclusions per Grade and Content Level for Each Webb Alignment Indicator

Content Area and Grade	2009 Form 1 (Spring)				2009 Form 2 (Summer)			
	Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
English II	Fully aligned	Partially aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned
Algebra I	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Weakly aligned	Fully aligned	Fully aligned
Biology	Fully aligned	Weakly aligned	Partially aligned	Fully aligned	Fully aligned	Weakly aligned	Partially aligned	Fully aligned

Recommendations

English II

1. **Increase DOK assessed by items included on the English II Form 1 test form relative to the Missouri CLEs (DOK consistency).** Panelists' ratings of item DOK for the 2009 English II EOC, Form 1 show that many items (52%) assess student knowledge at a lower level of cognitive complexity for the Reading strand than required by the content expectations. Improving alignment can be accomplished by modifying the language of existing items or by replacing items entirely. In either case, higher DOK for as few as three items covering Reading would increase alignment above the minimum criterion (from 48% of items to approximately 56%).
2. **Review the content emphasis on the English II assessments relative to the Missouri CLEs to ensure that the current emphasis corresponds with DESE's intentions.** Related to this point, we recommend that DESE consider the breadth of content covered within strands (i.e., number of CLEs matched to items). While the outcomes on the range-of-knowledge correspondence and balance-of-knowledge representation measures surpassed the Webb minimum criteria, certain CLEs received much greater content emphasis on the assessments. These findings suggest that the assessments may not cover the full range of the standards sufficiently. However, DESE may have intentionally selected certain CLEs for greater emphasis. If this was the case, this decision should be noted and explained in test documentation and in reports. Further, if a review of content emphasis occurs, Recommendation 1 above should be considered simultaneously.

Algebra I

1. **Increase DOK assessed by items included on the Algebra I Form 2 test form relative to the Missouri CLEs (DOK consistency).** Panelists' ratings of item DOK for 2009 Algebra I Form 2 show that it assesses students at a lower level of cognitive complexity than expected in the corresponding CLEs for the strands Numbers and Operations and Algebraic Relationships. Additionally, the assessment just met the criteria of 50% relative to the Data and Probability strand. Thus, over half of items require students to demonstrate content mastery using very basic cognitive skills (i.e., simple recall, low-level problem solving). As noted for English II, improving DOK alignment may involve minor item edits of stems and/or response options; or, some items could be replaced entirely. To increase alignment above the minimum criterion, approximately four items could be altered for the Number and Operations strand; three items for the Algebraic Relationships strand; and, two items for the Data and Probability strand.

2. **Review the content emphasis on the Algebra I assessments relative to the Missouri CLEs to ensure that the current emphasis corresponds with DESE’s intentions.** As with the English II assessments, some Algebra I content received much greater emphasis on the test forms. This weighting should be reviewed to ensure it is targeted to DESE’s intentions.

Biology

1. **Review the breadth of content covered *within* the Scientific Inquiry strand for both 2009 test forms (range-of-knowledge correspondence).** Both assessments include a sufficient number of items per content strand (well above six items); in addition, the Biology test forms cover a number of CLEs under Living Organisms and Ecosystems. However, the strand Scientific Inquiry did not receive as much emphasis, as reflected in the small number of CLEs targeted for assessment (approximately 6 of 17). Part of the reason for this outcome may be attributed to the nature of Scientific Inquiry as more of a *process* strand. Often, states intend for this type of strand either to receive less emphasis on assessments or the strand to be targeted in addition to other primary content strands. In the latter case, alignment review panelists frequently find it difficult to match process strands (in addition to content strands). Thus, it may be the case that panelists “under-matched” Scientific Inquiry. We cannot confirm this type of conclusion, however, without further review of items by state content experts. Regardless, assessment coverage of Scientific Inquiry is rather limited.

One additional comment regarding the Science course-level expectations pertains to the number of CLEs available for assessment. As the number of specific content expectations increases, the ability of the assessment to cover the range of content expectations adequately decreases. Solutions often considered by other states include: (a) increasing assessment length (more items), (b) redistributing item counts (particularly if some content receives greater emphasis), or (c) reviewing the content expectations to determine if some standards (CLEs) can be merged, targeted for classroom or local assessment, or even deleted from the state standards document.

2. **Increase DOK assessed by items included on the Biology test forms relative to the Missouri CLEs (DOK consistency).** The preponderance of items on the Biology test forms covering the strands Living Organisms and Scientific Inquiry assess student knowledge at a low level of complexity relative to the CLEs. Most of the discrepancy comes from numerous items rated as DOK Level 1 relative to the corresponding standard.

We noted that panelists found that the majority of CLEs expect students to demonstrate content knowledge at DOK Level 1 or 2. We would expect a higher proportion of content expectations to require higher level processing (i.e., Level 3 - strategic thinking, prediction), particularly for Science content. DESE may wish to review the CLEs in addition to the test forms to determine

whether the content standards expect students to demonstrate comprehension and application of Biology concepts at a sufficient level of complexity.

**MISSOURI ASSESSMENT PROGRAM (EOC):
ALIGNMENT FORMS VALIDATION STUDY**

TABLE OF CONTENTS

Chapter 1: Introduction	1
Chapter 2: Alignment Study Design and Methodology	3
Alignment of Assessments and Standards on Content and Performance.....	3
<i>Webb Alignment Method</i>	3
<i>Panelists</i>	4
<i>Materials</i>	7
<i>Procedures</i>	7
Reporting Webb Alignment Results	8
<i>Inter-rater Agreement Results</i>	9
<i>Categorical Concurrence</i>	10
<i>DOK Consistency</i>	10
<i>Range-of-Knowledge</i>	10
<i>Balance-of-Knowledge Representation</i>	10
Chapter 3 Results: English II	13
Inter-rater Agreement Results.....	13
Webb Alignment Results.....	13
<i>Categorical Concurrence</i>	14
<i>DOK Consistency</i>	14
<i>Range-of-Knowledge</i>	15
<i>Balance-of-Knowledge Representation</i>	16
Summary and Discussion of Results on Webb Alignment Indicators.....	17
Chapter 4 Results: Algebra I	20
Inter-rater Agreement Results.....	20
Webb Alignment Results.....	20
<i>Categorical Concurrence</i>	21
<i>DOK Consistency</i>	21
<i>Range-of-Knowledge</i>	22
<i>Balance-of-Knowledge Representation</i>	23
Summary and Discussion of Results on Webb Alignment Indicators.....	24
Chapter 5 Results: Biology	26
Inter-rater Agreement Results.....	26
Webb Alignment Results.....	27
<i>Categorical Concurrence</i>	27
<i>DOK Consistency</i>	27
<i>Range-of-Knowledge</i>	28
<i>Balance-of-Knowledge Representation</i>	29
Summary and Discussion of Results on Webb Alignment Indicators.....	30

Chapter 6: Summary and Recommendations	32
Recommendations	33
<i>English II</i>	33
<i>Algebra I</i>	33
<i>Biology</i>	34
References	36

LIST OF TABLES

Table 1. Summary Alignment Conclusions per Grade and Content Level for Each Webb Alignment Indicator	vii
Table 2.1 Professional and Demographic Characteristics of EOC Panelists	6
Table 2.2 Characteristics of 2009 EOC Test Forms Reviewed	7
Table 3.1 Intraclass Correlation Coefficients on DOK Ratings for English II	13
Table 3.2. Pairwise Comparisons on Content Agreement Between Reviewers	13
Table 3.3 Summary of Categorical Concurrence Results, English II, 2009 Form 1 (Summer)	14
Table 3.4 Summary of Categorical Concurrence Results, English II, 2009 Form 2 (Spring)	14
Table 3.5 Summary of DOK Results, English II, 2009 Form 1 (Spring)	15
Table 3.6 Summary of DOK Results, English II, 2009 Form 2 (Summer)	15
Table 3.7. Number of Content Strands and CLEs Eligible for Assessment on English II 2009 Forms 1 (Spring) and 2 (Summer)	15
Table 3.8. Summary of Range-of-Knowledge Results, English II, 2009 Form 1 (Spring)	16
Table 3.9. Summary of Range-of-Knowledge Results, English II, 2009 Form 2 (Summer)	16
Table 3.10. Summary of Balance-of-Knowledge Results, English II, 2009 Form 1 (Spring)	16
Table 3.11. Summary of Balance-of-Knowledge Results, English II, 2009 Form 2 (Summer)	17
Table 3.12. Summary Alignment Outcomes per Webb Criterion for EOC English II Test Forms	19
Table 4.1 Intraclass Correlation Coefficients on DOK Ratings for Algebra I	20
Table 4.2. Pairwise Comparisons on Content Agreement Between Reviewers	20
Table 4.3 Summary of Categorical Concurrence Results, Algebra I, 2009 Form 1 (Spring)	21
Table 4.4 Summary of Categorical Concurrence Results, Algebra I, 2009 Form 2 (Summer)	21
Table 4.5. Summary of DOK Results, Algebra I, 2009 Form 1 (Spring)	22
Table 4.6 Summary of Depth-of-Knowledge Results, Algebra I, 2009 Form 2 (Summer)	22
Table 4.7. Number of Content Strands and CLEs Eligible for Assessment on the Algebra I Test Forms 1 (Spring) and 2 (Summer)	22

Table 4.8. Summary of Range-of-Knowledge Results, Algebra I, 2009 Form 1 (Spring)	23
Table 4.9. Summary of Range-of-Knowledge Results, Algebra I, 2009 Form 2 (Summer)	23
Table 4.10. Summary of Balance-of-Knowledge Results, Algebra I, 2009 Form 1 (Spring)	23
Table 4.11. Summary of Balance-of-Knowledge Results, Algebra I, 2009 Form 2 (Summer)	24
Table 4.12. Summary Alignment Outcomes per Webb Criterion for 2009 Algebra I Test Forms 1 (Spring) and 2 (Summer)	25
Table 5.1 Intraclass Correlation Coefficients on DOK Ratings for Biology	26
Table 5.2. Pairwise Comparisons on Content Agreement Between Reviewers	26
Table 5.3. Summary of Categorical Concurrence Results, Biology, 2009 Form 1 (Spring)	27
Table 5.4. Summary of Categorical Concurrence Results, Biology, 2009 Form 2 (Summer)	27
Table 5.5. Summary of DOK Results, Biology, 2009 Form 1 (Spring)	28
Table 5.6 Summary of Depth-of-Knowledge Results, Biology, 2009 Form 2 (Summer)	28
Table 5.7. Number of Content Strands and CLEs Eligible for Assessment on EOC 2009 Form 1 (Spring) and Form 2 (Summer)	28
Table 5.8. Summary of Range-of-Knowledge Results, Biology, 2009 Form 1 (Spring)	29
Table 5.9. Summary of Range-of-Knowledge Results, Biology, 2009 Form 2 (Summer)	29
Table 5.10. Summary of Balance-of-Knowledge Results, Biology, 2009 Form 1 (Spring)	30
Table 5.11. Summary of Balance-of-Knowledge Results, Biology, 2009 Form 2 (Summer)	30
Table 5.12. Summary Alignment Outcomes per Webb Criterion for Biology Test Forms 1 (Spring) and 2 (Summer)	31
Table 6.1. Summary Alignment Conclusions per Course for Each Webb Alignment Indicator	32

MISSOURI ASSESSMENT PROGRAM (EOC) ALIGNMENT FORMS VALIDATION STUDY: TECHNICAL REPORT

Chapter 1: Introduction

The Missouri Department of Elementary and Secondary Education (DESE) requested an external independent alignment study of the End-of-Course (EOC) assessments for English II, Algebra I, and Biology. Specifically, the study evaluated the alignment of the Spring 2009 and Summer 2009 EOC test forms for each subject to the Missouri Course-Level Expectations (CLEs)². Missouri uses the EOC tests in the federal and state accountability programs. DESE awarded the Human Resources Research Organization (HumRRO) the contract to conduct this alignment study, along with Dr. Norman Webb as subcontractor.

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. 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.

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, and procedures. Chapters 3 through 5 provide alignment results for each EOC content test (English II, Algebra I, and Biology respectively). Finally, 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 detail on the content alignment results per test form. 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.

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

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.

Alignment of Assessments and Standards on Content and Performance

The term *alignment* in this context refers to the degree of consistency evident in instruction and measurement of the state's academic content standards. School curricula should include appropriate content detailed by the state. Any documents developed to accompany the content standards (e.g., performance descriptors, test specifications, curriculum resources) must accurately represent the expectations. Assessments must measure only the content specified in the standards, and student scores generated from these assessments should adequately reflect student knowledge of the content standards. An alignment study evaluates the strength of any or all of these relationships.

In general, alignment evaluations for 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 to 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?
- Does the assessment accurately measure student knowledge of content standards?

Several methods of alignment exist. Most methods involve ratings of several aspects of the assessment items relative to the content standards. The ratings are analyzed statistically to determine the extent of alignment. HumRRO collaborated with Dr. Norman Webb, using his method (1997; 1999; 2005), to conduct the EOC alignment reviews.

Webb Alignment Method

The Webb alignment method was designed originally for use with standard large-scale assessments. Dr. Webb has researched and refined this method over time (e.g., Webb, 1997; 1999; 2005), and his approach is supported by the Council of Chief State School Officers (CCSSO).

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 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) measures 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 to which it is linked. Webb refers to his comparison as *Depth-of-Knowledge consistency*.

Another measure examines the **range-of-knowledge correspondence** between the assessment and content standards. The range-of-knowledge measure examines in detail the breadth of knowledge represented by test items. 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 content objectives assessed by items.

Finally, the **balance-of-knowledge representation** criterion focuses on content coverage in yet more detail. In this case, the number of items matched to the content objective does matter. The balance of representation determines whether the assessment measures the content objectives equitably within each standard. Based on Webb's method, items should be distributed evenly across the objectives per standard for good balance. The balance-of-knowledge representation is determined by calculating an index, or score, for each standard. Each standard should meet or surpass a minimum index level to demonstrate adequate balance.

Panelists

HumRRO convened panels of Missouri educators and national content experts to review the EOC test forms. These panelists included current and former teachers, administrators, and curriculum specialists or district coordinators. Each panel included in-state and out-of-state panelists.

HumRRO coordinated the three review panels with the assistance of DESE. HumRRO received district contact information from DESE, sent inquiries of interest

across the state, and selected panelists with final approval from DESE³. In an effort to balance panels appropriately, HumRRO considered several factors when selecting candidates in addition to level and quality of experience: (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.

³ DESE requested exclusion of candidates only if an individual had met a maximum number of hours and payment through the State. HumRRO opted to exclude individuals who participated in item development activities within the past two years relevant to the tests they would be reviewing to reduce bias.

Table 2.1 Professional and Demographic Characteristics of EOC Panelists

Professional Position	Number of Panelists											Gender	
	Missouri	Out-of-State	1 SE	2 Heart of MO	3 KC	4 NE	5 NW	6 SC	7 SW	8 STL	9 Central	M	F
English II	3	3										2	4
Teacher	3	3							2	1			3
Administrator													
Curric. Spec.													
Algebra I	4	3										1	6
Teacher	4			1	1								4
Admin													
Curric. Spec.		3											
Biology	4	3				1			1			3	4
Teacher	4	2	1	1		1						1	3
Admin													
Curric. Spec.		1											
									1				

Materials

Panelists evaluated the alignment of the EOC items with the Missouri CLEs. This section describes the CLEs reviewed, test form structure, and ratings forms and instructions used by panelists.

Test Forms. Panelists evaluated one Spring 2009 EOC (Form 1) and one Summer 2009 EOC (Form 2) per course. Table 2.2 lists the characteristics of these test forms per course. The test form review included only operational items (no field-test or anchor items). This report does not include any examples of items or references to specific item content due to test security.

Table 2.2 Characteristics of 2009 EOC Test Forms Reviewed

Course	2009 Form 1 (Spring) Total Field-Test Items	2009 Form 2 (Summer) Total Field-Test Items
English II	36	36
Algebra I	36	36
Biology	47	46

Rating Forms and Instructions. Panelists rated the CLEs and test items using the electronic Webb Alignment Tool (WAT). These ratings included: (a) DOK ratings of Missouri CLEs 2.0, (b) DOK ratings of individual test items, and (c) content match of individual items to CLEs. Panelists received instruction sheets listing the rating tasks and forms. Appendix E presents examples of rating forms and instructions.

Procedures

Dr. Norman Webb directed this alignment review at the Assessment Resource Center (ARC), at the University of Missouri, Columbia, on July 15 and 16, 2009. While panels were convened in facilities procured through DESE, Dr. Webb directed the actual reviews independently of DESE.

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. Dr. Webb and his staff 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 leader per group from outside the state of Missouri facilitated the alignment process. English II had included six reviewers, while Algebra I and Biology each contained seven reviewers.

Within their small groups, designated leaders experienced with alignment studies and the WAT 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 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 the WAT on-line where they made their 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 sample items. After completing sample items, panelists rated each 2009 item using the WAT (starting with Form 1, followed by Form 2). Panelists assigned a primary CLE to an item based on their judgment that an item clearly measured this content. Reviewers could assign up to two additional CLEs if they believed an item assessed another standard equally to the primary standard. Panelists completed item ratings individually; however, group leaders led panelists through an adjudication process after reviewers completed all items; only the highly discrepant ratings were included in the adjudication process. During the adjudication process, panelists were not required to reach consensus.

All panelists finished tasks in approximately two days, although they completed their ratings at different times. At the end of the alignment review, panelists provided summary comments about the alignment study (Dr. Webb) and completed two types of surveys: (a) feedback survey on alignment training and process (Webb/HumRRO) and (b) hotel accommodations survey (DESE).

Reporting Webb Alignment Results

To reduce unnecessary repetition, this section presents a summary, or description, of the analyses conducted as part of the alignment review for each subject. The analyses are (a) interrater agreement and (b) summary results of the four Webb alignment indicators. These summary alignment outcomes and conclusions are based on the detailed numeric results produced by the WAT. The detailed numeric results for each subject are found in the appendices. The WAT software automatically generates these results after completion of the alignment review; HumRRO reviewed these results for accuracy. We present detailed results by subject in the next three chapters for English II, Algebra I, and Biology, respectively.

All of Webb's measures begin with calculations for each reviewer and progress to a summary of results across reviewers per content strand. First, we calculated the mean ratings across items for each panelist, and then we determined the mean rating across

panelists per strand. Results generally are presented at the *strand* level (i.e., Reading, Writing).

Inter-rater Agreement Results

In this subsection, we report on two types of agreement analyses concerning panelists' ratings. Panelists 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, N. M., 2005; Tinsley & Weiss, 1975).

For item DOK ratings, the WAT applies the ICC (C, k) statistic, which refers to the intraclass correlation (ICC) coefficient. 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

When evaluating agreement between categorical ratings such as CLE content match to items, a different form of agreement statistic is required. Several agreement measures exist to analyze categorical ratings (see Gwet, 2001; Webb, N. L., 2005). For these data, the WAT calculates a measure developed by Webb, which basically is an estimate of 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 then are averaged across reviewers per test form. Webb's decision criteria for pairwise comparisons are comparable to those for the ICC, although slightly less stringent for exact agreement results 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

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.

Categorical Concurrence

Categorical concurrence describes the extent to which the 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.

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, we asked panelists to separately rate the CLEs and the test items. Webb includes an alignment indicator that directly compares panelists' DOK ratings of content standards and test items, which he refers to as *depth-of-knowledge consistency*.

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

- 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.

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 strand⁵. According to Webb, the minimum acceptable index for a single strand is 0.70 (on a 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 CLEs⁶.

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.

⁵ 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>.

⁶ The balance results must be interpreted within the context of the range-of-knowledge representation findings. Calculations of the balance index only include those standards matched to items by reviewers rather than the full pool of standards available for assessment.

Chapter 3 Results: English II

In this chapter, we report the results of the alignment review for English II which include: (a) inter-rater agreement and (b) summary results on the four Webb alignment indicators. At the end of this chapter, we highlight and discuss key outcomes.

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 this statistic and decision criteria. Table 3.1 presents inter-rater agreement outcomes (ICC) for item DOK ratings. These results are listed separately for Test Forms 1 and 2. The ICC (C, k) results in Table 3.1 indicate the reviewers frequently applied the same DOK ratings to the same items. All ICCs indicate 'Good agreement' between reviewers.

Table 3.1 Intraclass Correlation Coefficients on DOK Ratings for English II

Type of Agreement	DOK Agreement Results for 2009 Form 1	DOK Agreement Results for 2009 Form 2
ICC	0.94	0.88

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 strand, substrand, and CLE level. The second correlation presented for each form displays results for partial agreement, reflecting an assessment of agreement between reviewers at only the strand level.

Table 3.2. Pairwise Comparisons on Content Agreement Between Reviewers

Course	Pairwise Comparisons on 2009 Form 1 (Spring)		Pairwise Comparisons on 2009 Form 2 (Summer)	
	Exact Content Match (Strand, Substrand, CLE)	Partial Content Match (Strand only)	Exact Content Match (Strand, Substrand, CLE)	Partial Content Match (Strand only)
English II	0.95	0.97	0.82	1.00

Webb Alignment Results

This section reviews the general outcomes of item analyses on the four Webb alignment indicators. These summary alignment outcomes and conclusions are based on the detailed numeric results produced by the WAT. The detailed numeric results for English II can be found in Appendix A.

Categorical Concurrence

Tables 3.3 through 3.4 summarize the English II alignment results on categorical concurrence for test forms reviewed. Table 3.3 shows that Form 1 (Spring) and Form 2 (Summer) each included at least six items assessing the Reading (M=30.17 items) and the Writing (M=17 items) strands. Thus, both assessment forms met the minimum alignment criterion on categorical concurrence.

Table 3.3 Summary of Categorical Concurrence Results, English II, 2009 Form 1 (Summer)

Course	Mean Number of Items per Strand for 2009 Form 1		Strands with at Least Six Items
	Reading	Writing	
English II	30.17	17	2 of 2

Table 3.4 Summary of Categorical Concurrence Results, English II, 2009 Form 2 (Spring)

Course	Mean Number of Items per Strand for 2009 Form 2		Strands with at Least Six Items
	Reading	Writing	
English II	30.17	17	2 of 2

DOK Consistency

Tables 3.5 and 3.6 summarize the DOK consistency results for the English II test forms. Most results from the DOK consistency analysis suggest that the English II test forms assess student knowledge at a comparable level of complexity as expected in the corresponding CLEs. Writing items in particular match the CLEs well. Items assessing Reading on Form 1 reveal an exception to this pattern—reviewers rated only 49% of items as matched to the cognitive level of the corresponding CLEs. As a result, Form 1 does not meet the minimum DOK consistency required to cover Reading.

Table 3.5 Summary of DOK Results, English II, 2009 Form 1 (Spring)

Course	Percent of Items with DOK At and Above the Level of the CLEs per Strand		Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Reading	Writing		
English II	49	87	1	Reading

Table 3.6 Summary of DOK Results, English II, 2009 Form 2 (Summer)

Course	Percent of Items with DOK At and Above the Level of the CLEs per Strand		Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Reading	Writing		
English II	59	87	2	0

Range-of-Knowledge

Table 3.7 lists the number of strands, substrands, and CLEs found in the Missouri CLEs compared with the number of items per test form. This table includes only CLEs assessed on the EOC test; additional locally assessed standards are not included in these counts.

Table 3.7. Number of Content Strands and CLEs Eligible for Assessment on English II 2009 Forms 1 (Spring) and 2 (Summer)

Number of Content Strands	Number of Substrands	Number of CLEs Available for Assessment	Total Items for Form 1	Total Items for Form 2
2	6	16	36	36

Tables 3.8 and 3.9 summarize the range-of-knowledge results for each test form produced by the WAT software. At least 50% of CLEs per strand should be assessed by one or more items for adequate coverage. Tables 3.8 and 3.9 reveal that reviewers matched over half of the CLEs per content strand to at least one item for both test forms. However, the number of CLEs assessed for Writing is just above the minimum criterion.

Table 3.8. Summary of Range-of-Knowledge Results, English II, 2009 Form 1 (Spring)

Course	Percent of CLEs per Strand Assessed by At Least One Item on 2009 Form 1		Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Reading	Writing		
English II	98	57	2 of 2	0

Table 3.9. Summary of Range-of-Knowledge Results, English II, 2009 Form 2 (Summer)

Course	Percent of CLEs per Strand Assessed by At Least One Item on 2009 Form 2		Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Reading	Writing		
English II	98	57	2 of 2	0

A list of all CLEs matched to items by panelists is presented in Appendix A.

Balance-of-Knowledge Representation

Tables 3.10 and 3.11 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 can be seen, both strands in both forms had indices higher than 0.70.

Table 3.10. Summary of Balance-of-Knowledge Results, English II, 2009 Form 1 (Spring)

Course	Balance Index per Strand for 2009 Form 1		Strands with Adequate Balance	Strands with Limited Balance
	Reading	Writing		
English II	0.79	0.96	2 of 2	0

Table 3.11. Summary of Balance-of-Knowledge Results, English II, 2009 Form 2 (Summer)

Course	Balance Index per Strand for 2009 Form 2		Strands with Adequate Balance	Strands with Limited Balance
	Reading	Writing		
English II	0.75	0.96	2 of 2	0

While results on the distribution of items indicate that the assessments met the minimum decision criterion ($M=0.70$ or higher), a closer examination of item distribution per CLE suggests that some CLEs receive much more emphasis than others. For example, Table A-7 in Appendix A with Form 1 results shows that the CLE R.3.C.1 (Text Structure - Use details from informational and persuasive text(s)...) received twice as much emphasis (Sum = 8 items matched by all reviewers) as any other CLE. In comparison, panelists determined that the CLEs R.1.I.1 (Making Connections...) and R.2.A.1 (Text Features - Analyze and evaluate the text features in grade-level text) each matched to only one item. In comparison, Table A-8 with Form 2 results shows a different pattern, but one that still reflects disproportionate content emphasis (i.e., Writing items assessed 4 of 8 CLEs; 3 of 4 assessed CLEs correspond to single items).

Summary and Discussion of Results on Webb Alignment Indicators

The content alignment review of the EOC assessments evaluated the Spring and Summer forms of the 2009 English II assessment compared to the Missouri CLEs. Test forms for a given administration cycle should be representative of the full set of items in the pool and, thus, should align appropriately to the content expectations. Alignment of large-scale assessments to state content standards is a requirement of NCLB legislation.

Overall, the alignment results on the EOC test forms for English II suggest that each form demonstrates adequate alignment with the Missouri CLEs on breadth and depth. One exception pertains to DOK consistency for Form 1 where panelists' ratings indicate that half the items do not match the cognitive level of the corresponding CLEs. In addition, items target some content disproportionately, which may be an intentional feature of the test forms but should be explained.

Summary alignment judgments are based on Webb (2005). These summary judgments focus on the percentage of content strands represented well by the assessment. 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%).

Webb’s alignment method does not allow for a *single* judgment of overall alignment across the four alignment indicators. Instead, results reflect areas of strength and weakness in alignment. However, one can get a sense of overall alignment between the assessments and standards by looking holistically at all of the alignment indicators.

Table 3.12 presents the summary alignment outcomes on the EOC English II test forms based on the above scale. The table includes a summary judgment for each Webb alignment indicator per grade assessment, based on the percentage of strands that met the minimum alignment criteria. This summary table links to the bottom row of tables in Appendix A (Tables A-1 through A-12); thus, these summary judgments reflect a final evaluation of each grade assessment per Webb criteria *across* the strands.

As indicated by green highlighting (refer to Table 3.12), a number of outcomes point to strong content alignment of the EOC to the Missouri CLEs. Each form reviewed clearly includes a sufficient number of operational items to cover the major content categories (strands), as demonstrated by the outcomes on categorical concurrence. Panelists found that items matched a sufficient number of CLEs per strand, indicating that the assessment covers reasonable breadth of content. Furthermore, the balance-of-knowledge representation results suggest that items are reasonably distributed, at least across CLEs matched by panelists.

Two features of the test forms may warrant review. First, the DOK level of items assessing Reading for Form 1 in particular should be increased for approximately half of items to better match the CLEs, as noted by the conclusion of ‘partially aligned’ in Table 3.12. Second, while the range-of-knowledge correspondence outcomes produced a final judgment of ‘fully aligned’ based on the Webb minimum criterion, both test forms assessed a relatively narrow range of CLEs (impact is greater for Writing). This issue, in conjunction with the finding noted on page 14 of this report regarding item distribution, suggests that DESE may wish to review content emphasis on the assessment. A disproportionate emphasis of some content may be intended by DESE, which could be confirmed and justified by the test blueprint. However, for those CLEs matched to only one item, DESE may consider whether this number is sufficient to demonstrate accurate assessment of student knowledge of these content expectations.

Table 3.12. Summary Alignment Outcomes per Webb Criterion for EOC English II Test Forms

2009 Form 1				2009 Form 2			
Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
Categorical Concurrence	Depth-of- Knowledge Consistency	Range-of- Knowledge Correspondence	Balance-of- Knowledge Representation	Categorical Concurrence	Depth-of- Knowledge Consistency	Range-of- Knowledge Correspondence	Balance-of- Knowledge Representation
Fully aligned (100%)	Partially aligned (50%)	Fully aligned (100%)	Fully aligned (100)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100)

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

Chapter 4 Results: Algebra I

In this chapter, we report the results of the alignment review for Algebra I which include: (a) inter-rater agreement and (b) summary results on the four Webb alignment indicators. At the end of this chapter, we highlight and discuss key outcomes.

Inter-rater Agreement Results

For item DOK ratings, the WAT applies the ICC (C, k) statistic, which refers to the intraclass correlation (ICC) coefficient. Refer to Chapter 2 for an explanation of this statistic and decision criteria. Table 4.1 presents inter-rater agreement outcomes for item DOK ratings (ICC). These results are listed separately for Test Forms 1 and 2. As shown by these results, the reviewers frequently applied the same DOK ratings to the same items, resulting in ‘good agreement.’

Table 4.1 Intraclass Correlation Coefficients on DOK Ratings for Algebra I

Type of Agreement	DOK Agreement Results for 2009 Form 1	DOK Agreement Results for 2009 Form 2
ICC	0.91	0.88

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 at the strand, substrand, and CLE level. The second correlation presented for each form displays results for partial agreement, reflecting an assessment of agreement between reviewers at only the strand level.

Table 4.2. Pairwise Comparisons on Content Agreement Between Reviewers

Course	Pairwise Comparisons on 2009 Form 1 (Spring)		Pairwise Comparisons on 2009 Form 2 (Summer)	
	Exact Content Match (Strand, Substrand, CLE)	Partial Content Match (Strand only)	Exact Content Match (Strand, Substrand, CLE)	Partial Content Match (Strand only)
Algebra I	0.75	0.94	0.75	0.92

These results on pairwise comparisons for Forms 1 and 2 indicate that reviewers showed ‘good agreement’ on CLEs matched to items, even for exact matches on content strand, substrand, and CLEs.

Webb Alignment Results

In this section, we review the general outcomes of item analyses on the four Webb alignment indicators. We based these summary alignment outcomes and conclusions on the detailed numeric results produced by the WAT. The detailed numeric results for Algebra I can be found in Appendix B.

Categorical Concurrence

Tables 4.3 and 4.4 summarize the Algebra I alignment results on categorical concurrence for Forms 1 and 2. These results ($M > 6$ in each case) indicate that the EOC test forms adequately cover the breadth of the Algebra I content strands that students are expected to know across grade levels. Table 4.3 shows that Form 1 (Spring) and Form 2 (Summer) each included at least six items per content strand. Thus, both assessment forms met the minimum alignment criterion on categorical concurrence.

Table 4.3 Summary of Categorical Concurrence Results, Algebra I, 2009 Form 1 (Spring)

Course	Mean Number of Items per Strand for 2009 Form 1			Strands with at Least Six Items
	Numbers and Operations	Algebraic Relationships	Data and Probability	
Algebra I	8.29	25.57	7.57	3 of 3

Table 4.4 Summary of Categorical Concurrence Results, Algebra I, 2009 Form 2 (Summer)

Course	Mean Number of Items per Strand for 2009 Form 2			Strands with at Least Six Items
	Numbers and Operations	Algebraic Relationships	Data and Probability	
Algebra I	7.29	25.43	8.57	3 of 3

DOK Consistency

Tables 4.5 and 4.6 summarize the DOK consistency results for each grade level of the EOC. Panelists' ratings on DOK consistency for Algebra I point to inconsistency in the extent to which the assessments measure student knowledge appropriately when compared to the Missouri CLEs. Reviewers rated over 60% of Form 1 items as matched to the corresponding CLEs on cognitive complexity. In comparison, only 50% of items on Form 2 were rated as assessing the CLEs at the appropriate cognitive level. These results indicate that an insufficient number of items on Form 2 assess the strands Numbers and Operations and Algebraic Relationships adequately for acceptable DOK consistency, and only 50% of items assessing the Data and Probability strand correspond with the cognitive level of the CLEs.

Table 4.5. Summary of DOK Results, Algebra I, 2009 Form 1 (Spring)

Course	Percent of Items with DOK At and Above the Level of the CLEs per Strand			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Numbers and Operations	Algebraic Relationships	Data and Probability		
Algebra I	67	67	63	3	0

Table 4.6 Summary of Depth-of-Knowledge Results, Algebra I, 2009 Form 2 (Summer)

Course	Percent of Items with DOK At and Above the Level of the CLEs per Strand			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Numbers and Operations	Algebraic Relationships	Data and Probability		
Algebra I	41	48	50	1	<ul style="list-style-type: none"> • Numbers and Operations • Algebraic Relationships

Range-of-Knowledge

Table 4.7 lists the number of strands and CLEs found in the Missouri CLEs compared with the number of items per test form. This table includes only CLEs assessed on the EOC test; additional locally assessed standards are not included in these counts.

Table 4.7. Number of Content Strands and CLEs Eligible for Assessment on the Algebra I Test Forms 1 (Spring) and 2 (Summer)

Number of Content Strands	Number of CLEs Available for Assessment	Total Items for Form 1	Total Items for Form 2
3	17	36	36

Tables 4.8 and 4.9 summarize the range-of-knowledge results for each test form produced by the WAT software. At least 50% of CLEs per strand should be assessed by one or more items for adequate coverage. Tables 4.8 and 4.9 summarize the range-of-knowledge results for Algebra I per content strand. Results for both test forms indicate that items adequately covered a range of CLEs for each strand; thus, reviewers matched most of the CLEs to items.

Table 4.8. Summary of Range-of-Knowledge Results, Algebra I, 2009 Form 1 (Spring)

Course	Percent of CLEs per Strand Assessed by At Least One Item on 2009 Form 1			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Numbers and Operations	Algebraic Relationships	Data and Probability		
Algebra I	100	91	80	3 of 3	0

Table 4.9. Summary of Range-of-Knowledge Results, Algebra I, 2009 Form 2 (Summer)

Course	Percent of CLEs per Strand Assessed by At Least One Item on 2009 Form 1			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Numbers and Operations	Algebraic Relationships	Data and Probability		
Algebra I	100	90	82	3 of 3	0

A list of all CLEs matched to items by panelists is presented in Appendix B.

Balance-of-Knowledge Representation

Tables 4.10 and 4.11 summarize the results on balance-of-knowledge representation per test form. An index of 0.70 or higher indicates adequate distribution of items among assessed CLEs. Reviewers' ratings indicated adequate balance-of-knowledge results for all three Algebra I strands for both EOC test forms.

Table 4.10. Summary of Balance-of-Knowledge Results, Algebra I, 2009 Form 1 (Spring)

Course	Balance Index per Strand for 2009 Form 1			Strands with Adequate Balance	Strands with Limited Balance
	Numbers and Operations	Algebraic Relationships	Data and Probability		
Algebra I	0.92	0.78	0.86	3	0

Table 4.11. Summary of Balance-of-Knowledge Results, Algebra I, 2009 Form 2 (Summer)

Course	Balance Index per Strand for 2009 Form 1			Strands with Adequate Balance	Strands with Limited Balance
	Numbers and Operations	Algebraic Relationships	Data and Probability		
Algebra I	0.81	0.72	0.82	3	0

One caveat to note regarding the results on balance pertains to the content emphasis given to strands by both assessments. The Numbers and Operations and the Algebraic Relationships strands receive greater emphasis compared to the Data and Probability strand. In addition, reviewers found that only one to two items assessed some CLEs within the Data and Probability strand. The emphasis of some strands more than others on the assessments may be intentional be DESE.

Summary and Discussion of Results on Webb Alignment Indicators

The content alignment review of the Algebra I EOC evaluated Spring and Summer 2009 test forms compared to the Missouri CLEs. Test forms for a given administration cycle should be representative of the full set of items in the pool and, thus, should align appropriately to the content expectations. Alignment of large-scale assessments to state content standards is a requirement of NCLB legislation.

The overall alignment results for the Algebra I test forms suggest that test items align well to the CLEs on breadth of content coverage. However, items on Form 2 (Summer) do not meet the DOK requirements of the CLEs. Specifically, panelists found that less than 50% of items assessed students at the same cognitive levels expected for the CLEs under the Numbers and Operations and Algebraic Relationships strands. It also should be noted that exactly 50% of items met depth requirements of the Data and Probability strand.

Summary alignment judgments are based on Webb's indicators (2005). These summary judgments focus on the percentage of content strands represented well by the assessment. 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%); and
- Weakly aligned – assessments align few strands (below 50%).

Webb's method does not allow for a *single* judgment of overall alignment across the four alignment indicators. Instead, results reflect areas of strength and weakness in

alignment. However, one can get a sense of overall alignment between the assessments and standards by looking holistically at all of the alignment indicators.

Table 4.12 presents the summary alignment outcomes for Algebra I based on the above scale. The table includes a summary judgment for each Webb alignment indicator per grade test form, based on the percentage of strands that met the minimum alignment criteria. This summary table links to the bottom row of each table in Appendix B (Tables B-1 through B-12); thus, these summary judgments reflect a final evaluation of each grade assessment per Webb criteria *across* the strands.

Table 4.12. Summary Alignment Outcomes per Webb Criterion for 2009 Algebra I Test Forms 1 (Spring) and 2 (Summer)

2009 Form 1				2009 Form 2			
Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
Categorical Concurrence	Depth-of- Knowledge Consistency	Range-of- Knowledge Correspondence	Balance-of- Knowledge Representation	Categorical Concurrence	Depth-of- Knowledge Consistency	Range-of- Knowledge Correspondence	Balance-of- Knowledge Representation
Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Fully aligned (100%)	Weakly Aligned (33%)	Fully aligned (100%)	Fully aligned (100%)

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

This chapter reports on the results of the alignment review for Biology which include: (a) inter-rater agreement and (b) summary results on the four Webb alignment indicators. At the end of this chapter, we highlight and discuss key outcomes. Key outcomes are highlighted and discussed at the end of this chapter.

Inter-rater Agreement Results

For item DOK ratings, the WAT applies the ICC (C, k) statistic, which refers to the intraclass correlation (ICC) coefficient. Refer to Chapter 2 for an explanation of this statistic and decision criteria. Table 5.1 presents inter-rater outcomes for item DOK ratings (ICC). These results are listed separately for Test Forms 1 and 2. As can be seen from the results, panelists reached “good agreement” on their DOK ratings for both forms.

Table 5.1 Intraclass Correlation Coefficients on DOK Ratings for Biology

Type of Agreement	DOK Agreement Results for 2009 Form 1	DOK Agreement Results for 2009 Form 2
ICC	0.94	0.94

Pairwise comparisons are used to evaluate agreement between categorical ratings such as CLE content match to items. For these data, the WAT calculates a measure developed by Norman Webb, which basically is an estimate of percent agreement between reviewers⁷. 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 at the strand, substrand, and CLE level. The second correlation presented for each form displays results for partial agreement, reflecting an assessment of agreement between reviewers at only the strand level.

Table 5.2. Pairwise Comparisons on Content Agreement Between Reviewers

Course	Pairwise Comparisons on 2009 Form 1 (Spring)		Pairwise Comparisons on 2009 Form 2 (Summer)	
	Exact Content Match (Strand, Substrand, CLE)	Partial Content Match (Strand only)	Exact Content Match (Strand, Substrand, CLE)	Partial Content Match (Strand only)
Biology	0.75	0.98	0.73	1.00

⁷ Refer to *Webb, N. L. (2005). Webb Alignment Tool (WAT): Training Manual* for a detailed discussion of the agreement analysis based on pair-wise comparisons.

Results of the pairwise comparisons indicate that reviewers showed ‘agreement on CLEs matched to items, even for exact matches on content strand, substrand, and CLEs.

Webb Alignment Results

In this section, we review the general outcomes of item analyses on the four Webb alignment indicators. We based these summary alignment outcomes and conclusions on the detailed numeric results produced by the WAT. The detailed numeric results for Biology can be found in Appendix C.

Categorical Concurrence

Tables 5.3 and 5.4 summarize the Biology alignment results on categorical concurrence for the 2009 forms. These results indicate that both EOC test forms adequately cover each Biology content strand with a sufficient number of items.

Table 5.3. Summary of Categorical Concurrence Results, Biology, 2009 Form 1 (Spring)

Course	Mean Number of Items per Strand for 2009 Form 1			Strands with at Least Six Items
	Living Organisms	Ecosystems	Scientific Inquiry	
Biology	22.14	12.57	20.43	3

Table 5.4. Summary of Categorical Concurrence Results, Biology, 2009 Form 2 (Summer)

Course	Mean Number of Items per Strand for 2009 Form 2			Strands with at Least Six Items
	Living Organisms	Ecosystems	Scientific Inquiry	
Biology	20.86	14.29	20	3

DOK Consistency

Tables 5.5 and 5.6 summarize the DOK consistency results for each Biology test form. The results indicate that both test forms sufficiently assess students at the appropriate DOK level for the strand covering Ecosystems by surpassing the minimum criterion. In comparison, over half of items targeting the CLEs for Living Organisms and Scientific Inquiry assess students at a lower cognitive level than expected. DESE may wish to review the items targeting these strands to determine if items require modification to meet the Biology CLEs.

Table 5.5. Summary of DOK Results, Biology, 2009 Form 1 (Spring)

Course	Percent of Items with DOK At or Above the Level of the CLEs per Strand			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Living Organisms	Ecosystems	Scientific Inquiry		
Biology	48	56	46	1	<ul style="list-style-type: none"> • Living Organisms • Scientific Inquiry

Table 5.6 Summary of Depth-of-Knowledge Results, Biology, 2009 Form 2 (Summer)

Course	Percent of Items with DOK At or Above the Level of the CLEs per Strand			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Living Organisms	Ecosystems	Scientific Inquiry		
Biology	48	70	31	1	<ul style="list-style-type: none"> • Living Organisms • Scientific Inquiry

Range-of-Knowledge

Table 5.7 lists the number of strands and CLEs found in the Missouri CLEs compared with the number of items per test form. This table includes only CLEs assessed on the Biology test forms; additional locally assessed standards are not included in these counts.

Table 5.7. Number of Content Strands and CLEs Eligible for Assessment on EOC 2009 Form 1 (Spring) and Form 2 (Summer)

Number of Content Strands	Number of CLEs Available for Assessment	Total Items for Form 1	Total Items for Form 2
3	40	47	46

Tables 5.8 and 5.9 summarize the range-of-knowledge results for each test form produced by the WAT software. At least 50% of CLEs per strand should be assessed by one or more items for adequate coverage. Results show that both forms target a sufficient range of CLEs for the strands Living Organisms and Ecosystems. In comparison, the assessments target a narrow range of CLEs covering Scientific Inquiry. While reviewers matched approximately 20 items to Scientific Inquiry on average on both forms, these items covered only about six of the 15 CLEs.

Table 5.8. Summary of Range-of-Knowledge Results, Biology, 2009 Form 1 (Spring)

Course	Percent of CLEs per Strand Assessed by At Least One Item on 2009 Form 1			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Living Organisms	Ecosystems	Scientific Inquiry		
Biology	86	94	42	2	• Scientific Inquiry

Table 5.9. Summary of Range-of-Knowledge Results, Biology, 2009 Form 2 (Summer)

Course	Percent of CLEs per Strand Assessed by At Least One Item on 2009 Form 2			Number of Strands Assessed Adequately	Specific Strands Not Assessed Adequately
	Living Organisms	Ecosystems	Scientific Inquiry		
Biology	86	94	42	2	• Scientific Inquiry

A list of all CLEs, including those matched to items by panelists, is presented in Appendix C.

Balance-of-Knowledge Representation

The fourth measure of alignment included in the Webb method is *balance-of-knowledge representation*, which describes the distribution of items linked to each CLE within each strand. Tables 5.10 and 5.11 summarize the results on balance-of-knowledge representation for each test form. These results suggest that the Biology test forms display adequate balance for all three Science strands. However, this is an example of a circumstance when the balance-of-knowledge outcomes should be interpreted with caution, particularly for the Scientific Inquiry strand. Recall from Tables 5.8 and 5.9 that, while there are a sufficient number of test items assessing Scientific Inquiry, these items target a small number of CLEs. Thus, the balance-of-knowledge outcomes below reflect item distribution only *among the CLEs actually assessed*. In addition, the findings on DOK consistency “trump” the balance results because items assess many of these CLEs at a lower level of complexity.

Table 5.10. Summary of Balance-of-Knowledge Results, Biology, 2009 Form 1 (Spring)

Course	Balance Index per Strand for 2009 Form 1			Strands with Adequate Balance	Strands with Limited Balance
	Living Organisms	Ecosystems	Scientific Inquiry		
Biology	0.81	0.81	0.83	3	0

Table 5.11. Summary of Balance-of-Knowledge Results, Biology, 2009 Form 2 (Summer)

Course	Balance Index per Strand for 2009 Form 2			Strands with Adequate Balance	Strands with Limited Balance
	Living Organisms	Ecosystems	Scientific Inquiry		
Biology	0.81	0.77	0.85	3	0

Summary and Discussion of Results on Webb Alignment Indicators

The content alignment review of Biology EOC assessments evaluated the Spring and Summer 2009 test forms compared to the Missouri CLEs. A test form for a given yearly administration should be representative of the full set of items in the pool and, thus, should align appropriately to the content expectations. Alignment of large-scale assessments to state content standards is a requirement of NCLB legislation.

The overall alignment results for the Biology test forms were mixed. The 2009 test forms include a sufficient number of items to adequately cover the breadth of the Science content strands for Biology. However, items target a narrow range of CLEs, especially for Scientific Inquiry. In addition, many items (over half for two of three strands) assess student knowledge of the content at a lower level of cognitive depth than required by corresponding CLEs. As a result, the findings on balance-of-knowledge representation should be interpreted with caution.

Summary alignment judgments are based on Webb (2005). These summary judgments focus on the percentage of content strands represented well by the assessment. 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%); and
- Weakly aligned – assessments align to less than half the strands (below 50%).

Webb's method does not allow for a *single* judgment of overall alignment across the four alignment indicators. Instead, results reflect areas of strength and weakness in alignment. However, one can get a sense of overall alignment between the assessments and standards by looking holistically at all of the alignment indicators.

Table 5.12 presents the summary alignment outcomes for Biology based on the above scale. The table includes a summary judgment for each Webb alignment indicator per grade assessment, based on the percentage of strands that met the minimum alignment criteria. This summary table links to the bottom row of each table in Appendix C (Tables C-1 through C-8); thus, these summary judgments reflect a final evaluation of each grade assessment per Webb criteria *across* the strands.

Table 5.12. Summary Alignment Outcomes per Webb Criterion for Biology Test Forms 1 (Spring) and 2 (Summer)

2009 Form 1 (Spring)				2009 Form 2 (Summer)			
Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Fully aligned (100%)	Weakly aligned (33%)	Partially aligned (67%)	Fully aligned (100%)	Fully aligned (100%)	Weakly aligned (33%)	Partially aligned (67%)	Fully aligned (100%)

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, along with Dr. Norman Webb, conducted a review of the EOC tests for English II, Algebra I, and Biology to examine content alignment to the Missouri CLEs. Alignment of assessments and achievement standards to the state academic content standards is a requirement of NCLB legislation.

The extent of alignment to the Missouri CLEs varied per content area and test form. The English II and Algebra I test forms covered the breadth of the content expectations well overall, as demonstrated by alignment outcomes on the Webb indicators categorical concurrence, range-of-knowledge correspondence, and balance-of-knowledge representation. Some review of content emphasis may be warranted to ensure that the assessments appropriately reflect the weighting intended by DESE. Depth-of-knowledge consistency results for English II and Algebra I revealed less consistent assessment of the CLEs. While most results indicated adequate assessment of the student’s cognitive skill level, some review of item complexity should be considered for English II, Form 1 on items targeting Reading and for Algebra II, Form 2 across strands.

The Biology test forms definitely include a sufficient number of items to cover each content strand (categorical concurrence). However, the assessments target a narrow range of Biology CLEs (range-of-knowledge correspondence). Consequently, we emphasize that the positive outcomes on balance-of-knowledge representation, suggesting reasonable item distribution across CLEs, reflect the small number of CLEs assessed. In addition, the majority of items on both test forms assess student knowledge at a lower level of cognitive complexity than required by the CLEs (depth-of-knowledge consistency).

Table 6.1 provides summary alignment conclusions for each course assessment per Webb alignment indicator.

Table 6.1. Summary Alignment Conclusions per Course for Each Webb Alignment Indicator

Course	2009 Form 1				2009 Form 2			
	Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
Eng II	Fully aligned	Partially aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned
Algebra I	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Fully aligned	Weakly aligned	Fully aligned	Fully aligned
Biology	Fully aligned	Weakly aligned	Partially aligned	Fully aligned	Fully aligned	Weakly aligned	Partially aligned	Fully aligned

Based on these results, HumRRO offers several recommendations to Missouri 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 note that DESE, along with the test developer, should review the results and recommendations relative to the test blueprints to determine if some outcomes per EOC content test are justifiable, meaning the state intentionally chose to emphasize some strands and CLEs over others. In these cases, DESE should consider explicitly including these justifications in test documentation.

Recommendations

English II

1. **Increase DOK assessed by items included on the English II, Form 1 relative to the Missouri CLEs (DOK consistency).** Panelists' ratings of item DOK for the 2009 English II EOC, Form 1 show that many items (52%) assess student knowledge at a lower level of cognitive complexity for the Reading strand than required by the content expectations. Improving alignment can be accomplished by modifying the language of existing items or by replacing items entirely. In either case, higher DOK for as few as three items covering Reading would increase alignment above the minimum criterion (from 48% of items to approximately 56%).
2. **Review the content emphasis on the English II assessments relative to the Missouri CLEs to ensure the current emphasis corresponds with DESE's intentions.** Related to this point, we recommend that DESE consider the breadth of content covered within strands (i.e., number of CLEs matched to items). While the outcomes on the range-of-knowledge correspondence and balance-of-knowledge representation measures surpassed the Webb minimum criteria, certain CLEs received much greater content emphasis on the assessments. These findings suggest that the assessments may not sufficiently cover the full range of the standards. However, DESE may have intentionally selected certain CLEs for greater emphasis. If this is the case, this decision should be noted and explained in test documentation and in reports. Further, if a review of content emphasis occurs, Recommendation 1 above should be considered simultaneously.

Algebra I

1. **Increase DOK assessed by items included on the Algebra I, Form 2 relative to the Missouri CLEs (DOK consistency).** Panelist ratings of item DOK for 2009 Algebra I, Form 2 show that it assesses students at a lower level of cognitive complexity than expected in the corresponding CLEs for the strands Numbers and Operations and Algebraic Relationships. Additionally,

the assessment just met the criteria of 50% relative to the Data and Probability strand. Thus, over half of items require students to demonstrate content mastery using very basic cognitive skills (i.e., simple recall, low-level problem solving). As noted for English II, improving DOK alignment may involve minor item edits of stems and/or response options or some items could be replaced entirely. To increase alignment above the minimum criterion, approximately four items could be altered for the Numbers and Operations strand; three items for the Algebraic Relationships strand; and, two items for the Data and Probability strand.

2. **Review the content emphasis on the Algebra I assessments relative to the Missouri CLEs to ensure the current emphasis corresponds with DESE’s intentions.** As with the English II assessments, some Algebra I content received much greater emphasis on the test forms. This weighting should be reviewed to ensure it is targeted to DESE’s intentions.

Biology

1. **Review the breadth of content covered *within* the Scientific Inquiry strand for both 2009 test forms (range-of-knowledge correspondence).** Both Biology assessments include a sufficient number of items per content strand (well above 6 items); in addition, the Biology test forms cover a number of CLEs under the Living Organisms and Ecosystems strands. However, the strand Scientific Inquiry did not receive as much emphasis, as reflected in the small number of CLEs targeted for assessment (approximately six of 17). Part of the reason for this outcome may be attributed to the nature of Scientific Inquiry as more of a *process* strand. Often, states intend for this type of strand either to receive less emphasis on assessments or to be targeted in addition to other primary content strands. In the latter case, alignment review panelists frequently find it difficult to match process strands (in addition to content strands). Thus, it may be the case that panelists “under-matched” Scientific Inquiry. We cannot confirm this type of conclusion, however, without further review of items by state content experts. Regardless, assessment coverage of Scientific Inquiry is rather limited.

One additional comment regarding the Science CLEs pertains to the number of CLEs available for assessment. As the number of specific content expectations increases, the ability of the assessment to cover the range of content expectations adequately decreases. Solutions often considered by other states include: (a) increasing assessment length (more items), (b) redistributing item counts (particularly if some content receives greater emphasis), or (c) reviewing the content expectations to determine if some standards (CLEs) can be merged, targeted for classroom or local assessment, or even deleted from the state standards document.

- 2. Increase DOK assessed by items included on the Biology test forms relative to the Missouri CLEs (DOK consistency).** The preponderance of items on the Biology test forms covering the strands Living Organisms and Scientific Inquiry assess student knowledge at a low level of complexity relative to the CLEs. Most of the discrepancy comes from numerous items rated as DOK Level 1 relative to the corresponding standard.

We noted, however, that panelists found that the majority of CLEs expect students to demonstrate content knowledge at DOK Level 1 or 2. We would expect a higher proportion of content expectations to require higher level processing (i.e., Level 3 - strategic thinking, prediction), particularly for Science content. DESE may wish to review the CLEs in addition to the test forms to determine whether the content standards expect students to demonstrate comprehension and application of Biology concepts at a sufficient level of complexity.

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
- No Child Left Behind Act of 2001. Public Law 107-110.
- Putka, D. & Sackett, P. (in press). *Reliability and validity*.
- 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).

