

**Missouri Assessment Program (MAP)
Alignment Forms Validation Study:
Technical Report**

**Leslie R. Taylor
Hilary L. Campbell
Rebecca Dvorak
Richard C. Deatz
Lisa E. Koger
Milton 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

February 19, 2010

Missouri Assessment Program (MAP) Alignment Forms Validation Study: Technical Report

**Leslie R. Taylor
Hilary L. Campbell
Rebecca Dvorak
Richard C. Deatz
Lisa E. Koger
Milton 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

February 19, 2010

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 (MAP) for Communication Arts-Reading and Writing, Mathematics, and Science. Specifically, the study evaluated the alignment of the MAP test forms (Grades 3 through 8 in Communication Arts and Math and Grades 5 and 8 in Science) to the Missouri Grade-Level Expectations¹. Missouri uses the MAP test in the federal and state accountability programs. DESE awarded 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 (NCLB) Act 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 that is 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 the state assessment system. Alignment results should demonstrate that the assessments represent the full range of the 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.

DESE has contracted with CTB/McGraw-Hill to produce and score the MAP tests. The MAP tests included the following three types of items: selected response, constructed response, and performance event/writing prompts. The MAP tests are to be aligned with the GLEs that are to be assessed on the state assessment and are to conform to the grade-level test blueprints. Student performance is reported using four achievement levels: Advanced, Proficient, Basic, and Below Basic.

Methodology

HumRRO convened six review panels of Missouri educators, as well as out-of-state content experts, to review the test forms by grade span (e.g., Grades 3, 4, and 5 or Grades 6, 7, and 8). HumRRO received district contact information from DESE, sent inquiries of interest across the state, and selected panelists with final approval from

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

DESE². In an effort to balance panels appropriately, HumRRO considered several factors when selecting candidates in addition to level and quality of experience: (1) region of origin in Missouri, (2) other demographic factors (e.g., rural/suburban, gender), and (3) status as a new or former panelist.

HumRRO used the alignment method developed by Norman Webb (1997; 1999; 2005) to evaluate the alignment of the 2010 and 2011 MAP test forms for Communication Arts, Mathematics, and Science to the Missouri Grade-Level Expectations. As part of this method, reviewers rate individual test items on the cognitive complexity and content assessed relative to the Missouri Grade-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, GLE) 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 consistency – compares the cognitive complexity ratings of the items with the complexity ratings of each content standard.

Summary of Results

Key Findings and Conclusions

The extent of alignment to the Missouri Grade-Level Expectations differed considerably per content area and grade. The 2010 and 2011 test forms for Mathematics demonstrated the strongest alignment to the GLEs. The Communication Arts test forms displayed the most variability in alignment across grades and alignment criteria. For example, Communication Arts results suggest that the majority of grade-level test forms assess students on a range of the GLEs within content strands/Big Ideas. Furthermore, the test forms assess the major Reading categories with a sufficient number of items for over half of the Big Ideas, although the Writing assessment may warrant review to ensure that these content expectations receive adequate emphasis as well. In comparison, the Communication Arts items tended to cluster around a small number of assessed GLEs, producing unbalanced content coverage. The test blueprints provide guidance on emphasis across strands within a content area and not at the GLE level. As a result, the number of items per strand may meet the test blueprint guidance, but still have unbalanced content coverage across GLEs within a strand. Finally, over

² 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.

50% of items on the majority of Communication Arts test forms assess students at a lower level of cognitive processing than required in the Missouri Grade-Level Expectations.

Findings for Science require some additional explanation for the following reasons. In matching items to Science GLEs, panel members matched items against the following choices: all strands, concepts, and assessed GLEs for the grade span (Grade Span 3-5 and Grade Span 6-8). As a result, when the panel members then matched the items to the GLEs, they could match an item to any of the grade span's assessed GLEs. For Grade 5, panel members matched the 63 items from the 2010 form and 64 items on the 2011 form to the 149 rated strands, concepts, and assessed GLEs for Grade Span 3-5. If panel members were restricted to matching against only those GLEs assessed for Grade 5, they would have been matching the items to only 56 GLEs. Grade 8 procedures were the same, but for Grade Span 6-8. Approximately 16% of the panelists' ratings matched items to Grade 3 standards, 22% to Grade 4 standards, and 62% to Grade 5 standards for the Grade 5 assessment. Approximately 24% of panelists' ratings matched items to Grade 6 standards, 27% to Grade 7 standards, and 50% to Grade 8 standards on the Grade 8 assessment. All panelists ratings are provided in the Appendices Tables C-13 and C-14.

The decision to have panelists match items to the grade span rather than to a single grade most directly affects the results in range of knowledge. For range of knowledge, only 1 strand was found to be adequately assessed from the possible 32 Science strands across both forms and grade levels. However, matching the 63 and 64 items for Grade 5 forms 2010 and 2011, respectively, to 56 rather than 149 choices provides a far higher likelihood for matching at least 50% of the GLEs within each strand. The same logic holds for matching the 65 and 64 items for Grade 8 forms 2010 and 2011, respectively, to 82 rather than 219 choices.

Categorical concurrence should not be affected since the strands remain constant across all grades.

Depth of knowledge may have been affected, but not to a large extent. The impact would probably be that more items were at the same or higher DOK level as the standards since the items also were being matched to lower grades' GLEs rather than only for Grades 5 and 8 (assuming that there was a tendency for standards to have lower DOK requirements at lower grades). This may have actually increased the number of strands that were determined to be adequately assessed.

Balance of knowledge also may be higher as a result of the items being dispersed over more choices. A quick scan of the data matches for Grades 5 and 8 found that the most frequently selected choices had only one or two items matched and very few choices with more than three items matched.

However, it is not possible to examine the data to match only with the Grade 5 and Grade 8 GLEs. Panelists matched items to standards from the grade span, and we

cannot ascertain what standards they would have matched, or if they would have matched standards at all, if only a single grade's standards had been presented. All analyses and results for science in this report were based on the assumption that each assessment represented a three-grade span.

Alignment of MAP Test Forms to Missouri Grade-Level Expectations

Table 1 provides summary conclusions on the alignment of the MAP to the Missouri Grade-Level Expectations for Communication Arts, Mathematics, and Science per grade tested. The conclusions are based on the following decision criteria (Webb, 2005):

- Fully aligned – assessments align to all content strands (100%) (indicated in green in Table 1);
- Highly aligned – assessments align to the majority of strands (70%–90%) (also indicated in green);
- Partially aligned – assessments align well to some strands (50%–69%) (yellow indicates a two thirds majority; orange, exactly 50%);
- Weakly aligned – assessments align to less than half the strands (below 50%) (indicated in red).

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

	2010 Test Form				2011 Test Form			
	Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
	CC	DOK	ROK	BOK	CC	DOK	ROK	BOK
Comm Arts								
3	High	Partial	High	Weak	Partial	Partial	High	Partial
4	Partial	Partial	Partial	Partial	Partial	Weak	Full	Partial
5	Partial	Partial	High	Partial	Partial	Partial	Partial	Partial
6	Partial	Weak	High	Weak	Partial	Partial	Full	Partial
7	Partial	Weak	Full	Partial	Partial	Partial	High	Weak
8	Partial	Partial	High	Partial	Partial	High	Full	Partial
Math								
3	Full	Full	Full	Full	Full	Full	Full	Full
4	Full	Full	Full	Full	Full	Full	Full	High
5	Full	High	Full	Full	Full	Full	Full	Full
6	Full	High	Full	High	Full	High	Full	High
7	Full	Full	Full	Partial	Full	Full	Full	Partial
8	Full	Full	Full	Full	Full	Partial	Full	Partial
Science								
5	High	Full	Weak	Full	Partial	High	Weak	Full
8	Partial	Partial	Weak	High	High	Weak	Weak	High

Note: CC = Categorical Concurrence; DOK = Depth-of-knowledge Consistency; ROK = Range-of-knowledge Correspondence; BOK = Balance-of-knowledge Representation

Recommendations

Communication Arts

1. Consider ways to increase overall content coverage on the assessments, particularly for Writing content expectations (categorical concurrence). The content expectations composing the Big Ideas on Writing-Process and Writing-Forms/Types that Missouri expects students to know currently appear underrepresented (fewer than 6 items per Big Idea) on the assessments, resulting in a conclusion of 'partial alignment' overall. Coverage of major content categories could be increased or explained in several ways: (a) increase number of items [approximately 4-6 selected response (SR) items per Big Idea], (b) if constructed response (CR) items target multiple content areas, provide more explicit description of possible content coverage for these items in test documentation to gain more transparency in

- demonstrating alignment, (c) consider developing, or modifying, CR items to target additional content, and/or (d) explicitly note in GLEs and test documentation why this Writing content is not assessed at the state level and describe how students are expected to demonstrate this knowledge in other ways.
- 2. Evaluate the cognitive complexity assessed by items relative to the Missouri Grade-Level Expectations for both grade-level test forms (depth-of-knowledge consistency).** With the exception of the 2011 form for Grade 8, the panelists reviewing these assessments rated a number of items as less demanding cognitively than the Missouri Grade-Level Expectations. Thus, the assessments may not adequately reflect the rigor of the state standards for some content expectations. This finding is not uncommon among large-scale assessments. However, such a circumstance also is not an inevitable consequence of standardized testing. The number of adjacent ratings (DOK of 1 vs. 2) given by reviewers suggests only moderate discrepancy between items and GLEs. Thus, increasing cognitive complexity may require minor modifications to items.
 - 3. Review the ratio of items assigned to assessed GLEs within each Big Idea for all grades to evaluate content emphasis on the assessment (balance-of-knowledge representation).** While the majority of grade-level Communication Arts test forms assessed a range of GLEs per strand, the distribution of items among these GLEs appears unbalanced. In other words, reviewer ratings suggest that a number of items cluster around one to two GLEs. This type of problem can be remedied in several ways: (a) increase the number of items assigned to GLEs with low emphasis, (b) redistribute more evenly the existing points (requiring some new item construction or modification) among GLEs, or (c) provide more explicit justification for uneven content emphasis. The solution chosen depends on various constraints (usually time and money) that exist for DESE and for the test vendor.

Mathematics

- 1. Review item assignment to content expectations for increased alignment to assessed GLEs within each strand for Grades 7 and 8 (balance-of-knowledge representation).** The majority of results for the grade-level Math test forms indicate high alignment to the GLEs. One area that DESE may wish to review is the item distribution among GLEs, particularly for assessment of the Number and Operations strand and the Data and Probability strand on the Grade 7 and 8 test forms. The Math test forms demonstrated alignment to a broad range of GLEs; thus, item clustering around several GLEs within these strands is the most likely explanation for reduced alignment in these cases. As with Communication Arts, three options may be considered for increasing balanced alignment (see above). The second option of redistributing points/items among GLEs may be the most practical option, even between strands, since about half of the total items

were matched by reviewers to GLEs within the Numbers and Operations and Algebraic Relationships strands.

Science

1. **Review the breadth of content covered on the 2011 test forms for Grades 5 and 8 to increase alignment to the Missouri Grade-Level expectations (categorical concurrence and range-of-knowledge correspondence).** The results on the test forms for Grades 5 and 8 indicate that these assessments do not meet the minimum criteria for several alignment measures when compared to the GLEs to be assessed for the grade spans (Grades 3-5 and Grades 6-8). The most critical issue pertains to the small percentage of Science GLEs within the grade span assessed by each grade-level test form, which is evident from the range-of-knowledge representation results. Thus, the assessments do not adequately “cover the full range of content specified in the State’s academic content standards” (USDE, 2004, p.41) for the grade span. If the state considers all of these content expectations for the grade span important for students to know in order to demonstrate mastery of grade span Science concepts, then the MAP should assess a larger proportion of the grade span content expectations.

This issue may be a result of combining, for this study, the GLEs for all grades within the grade spans (3-5 and 6-8) rather than examining the alignment of the Grades 5 and 8 Science assessment only to the that grade’s science GLEs. Examining only the grade-specific GLEs with the grade-level assessment provides a far higher likelihood of matching at least 50% of the GLEs within each strand to a test item to meet the Webb criterion for range of knowledge.

Related to content coverage at a broader level, reviewers found that some strands were assessed by fewer than six items³. Specifically, the 2011 Grade 5 test form and the 2010 Grade 8 test form did not meet the minimum criterion for adequate assessment of the strands Force and Motion, Universe, and Science and Technology. This outcome also is a symptom of an unbalanced ratio of test items to standards; however, the alignment issue is less critical for two reasons: (a) the mean number of items matched is very close to six in each case, and (b) the content emphasis on these assessments is comparable to the test blueprint. While some researchers argue that a minimum of six items is arbitrary, an assessment should include a sufficient number of items for accurate assessment of what students know to produce valid scores.

2. **Evaluate the cognitive complexity assessed by items relative to the Missouri Grade-Level Expectations on the Grade 8 test forms (depth-of-**

³ Since strands remained the same for across the grade spans, this criterion should not have been impacted by matching to all GLEs across the grade span.

knowledge consistency). Reviewer ratings of item DOK for Grades 5 and 8 suggest that the test forms assess a lower level of cognitive complexity overall than required by the content expectations⁴. The results for the Grade 8 test forms, in particular, indicate a more marked discrepancy (majority of items assessed as DOK level 1, while many GLEs rated as DOK level 2⁵). Two issues deserve consideration. First, while the magnitude of discrepancy between items and GLEs is low (DOK of 1 vs. 2), the number of items falling below the cognitive complexity level expected in corresponding GLEs is high. As a result, students rarely must demonstrate knowledge at the same level as the content standards. Second, and somewhat surprisingly, the performance expected of students in the majority of GLEs qualifies as lower-order cognitive processing. Relatively few GLEs expect students to master Science concepts at a higher level requiring complex reasoning (DOK level 3)⁶. Science concepts often involve greater *difficulty* due to the cumulative nature of Science knowledge acquisition. While difficulty and complex cognitive processing are correlated, *difficult* concepts requiring more prerequisite knowledge do not necessarily involve in-depth cognitive processing. DESE and the test developer may wish to review the GLEs in addition to the test forms to further examine whether the test items expect students to demonstrate comprehension and application of science concepts at the cognitive complexity level required of the students by the GLEs.

⁴ The results for this criterion should not have been adversely impacted by examining items to all GLEs within each grade span.

⁵ The HumRRO Alignment Panel rated 48 GLEs at DOK level 2 while the DESE Standards Writing Committee rated 49 GLEs at DOK level 2. There were three differences between the ratings by the Panel and the Committee. One GLE was rated one level higher by the Panel and 2 GLEs were rated one level higher by the Committee.

⁶ Both the Panel and the Committee rated 4 GLEs at DOK level 3 and 1 GLE at DOK level 4.

**MISSOURI ASSESSMENT PROGRAM (MAP):
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>	8
<i>Procedures</i>	9
Chapter 3 Results: Communication Arts	11
Inter-rater Agreement Results.....	11
Webb Alignment Results.....	13
<i>Categorical Concurrence</i>	14
<i>Depth-of-Knowledge Consistency</i>	15
<i>Range-of-Knowledge Correspondence</i>	16
<i>Balance-of-Knowledge Representation</i>	18
Summary and Discussion of Results on Webb Alignment Indicators.....	20
Chapter 4 Results: Mathematics	23
Inter-rater Agreement Results.....	23
Webb Alignment Results.....	25
<i>Categorical Concurrence</i>	25
<i>Depth-of-Knowledge Consistency</i>	26
<i>Range of Knowledge</i>	28
<i>Balance-of-Knowledge Representation</i>	30
Summary and Discussion of Results on Webb Alignment Indicators.....	32
Chapter 5 Results: Science	35
Inter-rater Agreement Results.....	35
Webb Alignment Results.....	37
<i>Categorical Concurrence</i>	38
<i>Depth-of-Knowledge Consistency</i>	39
<i>Range of Knowledge</i>	40
<i>Balance-of-Knowledge Representation</i>	42
Summary and Discussion of Results on Webb Alignment Indicators.....	44
Chapter 6 Summary and Recommendations	47
Recommendations	50
<i>Communication Arts</i>	50
<i>Mathematics</i>	51
<i>Science</i>	51
References	54

TABLE OF CONTENTS (CONTINUED)

List of Tables

Table 1. Summary Degree of Alignment Conclusions per Grade and Content Level for Each Webb Alignment Indicator	ix
Table 2.1 Professional and Demographic Characteristics of MAP Panelists.....	6
Table 2.2 Characteristics of 2010 and 2011 MAP Test Forms Reviewed	8
Table 3.1 Interclass Correlation Coefficients on DOK Ratings for Communication Arts	12
Table 3.2. Pair-Wise Comparisons on Content Agreement Between Communication Arts Reviewers	13
Table 3.3 Summary of Categorical Concurrence Results for Communication Arts-Reading and Writing per Big Idea, 2010 Test Form.....	14
Table 3.4 Summary of Categorical Concurrence Results for Communication Arts-Reading and Writing per Big Idea, 2011 Test Form.....	14
Table 3.5 Summary of Depth-of-Knowledge Results, Communication Arts, 2010 Test Form	16
Table 3.6 Summary of Depth-of-Knowledge Results, Communication Arts, 2011 Test Form	16
Table 3.7. Number of Content Strands and GLEs Eligible for Assessment on MAP Communication Arts 2010 and 2011 Test Forms	17
Table 3.8. Summary of Range-of-Knowledge Results, Communication Arts-Reading and Writing, 2010 Test Form	17
Table 3.9. Summary of Range-of-Knowledge Results, Communication Arts-Reading and Writing, 2011 Test Form	18
Table 3.10. Summary of Balance-of-Knowledge Results, Communication Arts-Reading and Writing, 2010 Test Form	19
Table 3.11. Summary of Balance-of-Knowledge Results, Communication Arts-Reading and Writing, 2011 Test Form	19
Table 3.12. Summary Degree of Alignment Outcomes per Webb Criterion for MAP Grade Level Tests in Communication Arts – Reading and Writing.....	21
Table 4.1 Interclass Correlation Coefficients on DOK Ratings for Math	24
Table 4.2. Pair-Wise Comparisons on Content Agreement for Math Reviewers	25
Table 4.3 Summary of Categorical Concurrence Results, Math, 2010 Test Form	26
Table 4.4 Summary of Categorical Concurrence Results, Math, 2011 Test Form	26
Table 4.5. Summary of Depth-of-Knowledge Results, Math, 2010 Test Form	27
Table 4.6 Summary of Depth-of-Knowledge Results, Math, 2011 Test Form	28
Table 4.7. Number of Content Strands and GLEs Eligible for Assessment on Math 2010 and 2011 MAP Test Forms.....	29
Table 4.8. Summary of Range-of-Knowledge Results, Math, 2010 Test Form	29
Table 4.9. Summary of Range-of-Knowledge Results, Math, 2011 Test Form	29
Table 4.10. Comparison of GLEs Matched to Items with GLEs Available for Assessment on Math 2010 and 2011 MAP Test Forms.....	30
Table 4.11. Summary of Balance-of-Knowledge Results, Math, 2010 Test Form	31

Table 4.12. Summary of Balance-of-Knowledge Results, Math, 2011 Test Form	31
Table 4.13. Summary Alignment Outcomes per Webb Criterion for MAP Grade Level Tests in Math	33
Table 5.1 Interclass Correlation Coefficients on DOK Ratings for Science	36
Table 5.2. Pair-Wise Comparisons on Content Agreement Between Reviewers for Science, 2010 and 2011 Test Forms.....	36
Table 5.3. Summary of Categorical Concurrence Results, Science, 2010 Test Form...	38
Table 5.4. Summary of Categorical Concurrence Results, Science, 2011 Test Form...	38
Table 5.5. Summary of Depth-of-Knowledge Results, Science, 2010 Test Form.....	40
Table 5.6 Summary of Depth-of-Knowledge Results, Science, 2011 Test Form.....	40
Table 5.7. Number of Content Strands and GLEs Eligible for Assessment on MAP Science 2010 and 2011 Test Forms.....	41
Table 5.8. Summary of Range-of-Knowledge Results, Science, 2010 Test Form.....	41
Table 5.9. Summary of Range-of-Knowledge Results, Science, 2011 Test Form.....	42
Table 5.10. Comparison of GLEs Matched to Items with GLEs Available for Assessment for 2010 and 2011 MAP Science Test Forms.....	42
Table 5.11. Summary of Balance-of-Knowledge Results, Science, 2010 Test Form	43
Table 5.12. Summary of Balance-of-Knowledge Results, Science, 2011 Test Form	44
Table 5.13. Summary Alignment Outcomes per Webb Criterion for MAP Science Tests	45
Table 6.1. Summary Degree of Alignment Conclusions per Grade and Content Level for Each Webb Alignment Indicator	49

List of Tables

Figure 1. Missouri Regional Professional Development Centers Map	7
--	---

MISSOURI ASSESSMENT PROGRAM (MAP) 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 Missouri Assessment Program (MAP) for Communication Arts, Mathematics, and Science. Specifically, the study evaluated the alignment of the MAP test forms (Grades 3 through 8 in Communication Arts and Math and Grades 5 and 8 in Science) to the Missouri Grade-Level Expectations⁷. Missouri uses the MAP test in the federal and state accountability programs. DESE awarded 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 (NCLB) Act 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 the state assessment system. Alignment results should demonstrate that the assessments represent the full range of the 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.

DESE has contracted with CTB/McGraw-Hill to produce and score the MAP tests. The MAP tests included the following three types of items: selected response, constructed response, and performance event/writing prompts. The MAP tests are to be aligned with the GLEs that are to be assessed on the state assessment and are to conform to the grade-level test blueprints. Student performance is reported using four achievement levels: Advanced, Proficient, Basic, and Below Basic.

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 content area. Finally, Chapter 6 provides recommendations for DESE to strengthen the alignment of the MAP over time.

⁷ Missouri Grade-Level Expectations can be found at <http://dese.mo.gov/divimprove/curriculum/GLE/>

Additional information is provided in the appendices of this report. Appendices A through C contain tables providing more detail on the content alignment results per grade-level 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.

Chapter 2 Alignment Study Design and Methodology

In this section, we discuss key concepts related to alignment research, followed by a description of the alignment evaluations and methods used as part of the Missouri MAP 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 must include appropriate content laid out by the state. Any documents developed to accompany the content standards (e.g., performance descriptors, test specifications, teaching guides) 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 used the alignment method developed by Norman Webb (1997; 1999; 2005) to evaluate the MAP.

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: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and 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 mathematics 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 this 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 looks in greater detail at 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 lay out 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 MAP test forms. These panelists included current and former teachers, administrators, and curriculum specialists/district coordinators. For each panel, the group consisted of five in-state Missouri panelists and two out-of-state panelists.

HumRRO developed six 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 DESE ⁸. In an effort to balance panels appropriately, HumRRO considered several factors when selecting candidates in addition to level and quality of experience: (1) region of origin in Missouri, (2) other demographic factors (e.g., rural/suburban, gender), and (3) status as a new or former panelist. Table 2.1 presents the characteristics of the panelists per grade level of the MAP.

⁸ 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 MAP Panelists

Professional Position	Number of Panelists		Special Certifications	Region of Origin in Missouri (based on Missouri RPDCs)											Gender		Ethnicity		
	Missouri	Out-of-State		1-SE	2-Heart	3-KC	4-NE	5-NW	6-SC	7-SW	8-STL	9-Central	10-Mo	11-Mo S	11-Mo W	M	F	White, Non-Hispanic	Black, Non-Hispanic
Comm Arts, Grades 3-5	5	2		0	0	0	0	1	1	1	2	0	0	0	2	5	7	0	0
Teacher	2	1	0	--	--	--	--	--	--	--	--	--	--	--					
Administrator Curriculum	2	0	0																
Specialist	1	1	1	--	--	--	--	--	--	--	--	--	--	--					
Comm Arts, Grades 6-8	5	2		1	0	1	1	0	0	1	0	0	0	1	2	5	4	2	1
Teacher	4	1	1	--	--	--	--	--	--	--	--	--	--	--					
Administrator Curriculum	0	1	0	--	--	--	--	--	--	--	--	--	--	--					
Specialist	1	0	1																
Math, Grades 3-5	5	2		0	1	1	1	0	1	1	0	0	0	0	2	5	7	0	0
Teacher	5	1	0	--	--	--	--	--	--	--	--	--	--	--					
Administrator Curriculum	0	0	0																
Specialist	0	1	1	--	--	--	--	--	--	--	--	--	--	--					
Math, Grades 6-8	5	2		0	1	1	0	1	1	1	0	0	0	0	2	5	7	0	0
Teacher	4	0	2																
Administrator Curriculum	0	1	0	--	--	--	--	--	--	--	--	--	--	--					
Specialist	1	1	1	--	--	--	--	--	--	--	--	--	--	--					
Science, Grade 5	5	2		0	2	0	0	1	0	1	1	0	0	0	1	6	7	0	0
Teacher	5	2	1	--	--	--	--	--	--	--	--	--	--	--					
Administrator Curriculum	0	0	0																
Specialist	0	0	0																
Science, Grade 8	5	2		0	1	1	0	2	0	0	1	0	0	0	2	5	7	0	0
Teacher	4	2	1	--	--	--	--	--	--	--	--	--	--	--					
Administrator Curriculum	0	0	0																
Specialist	1	0	1																

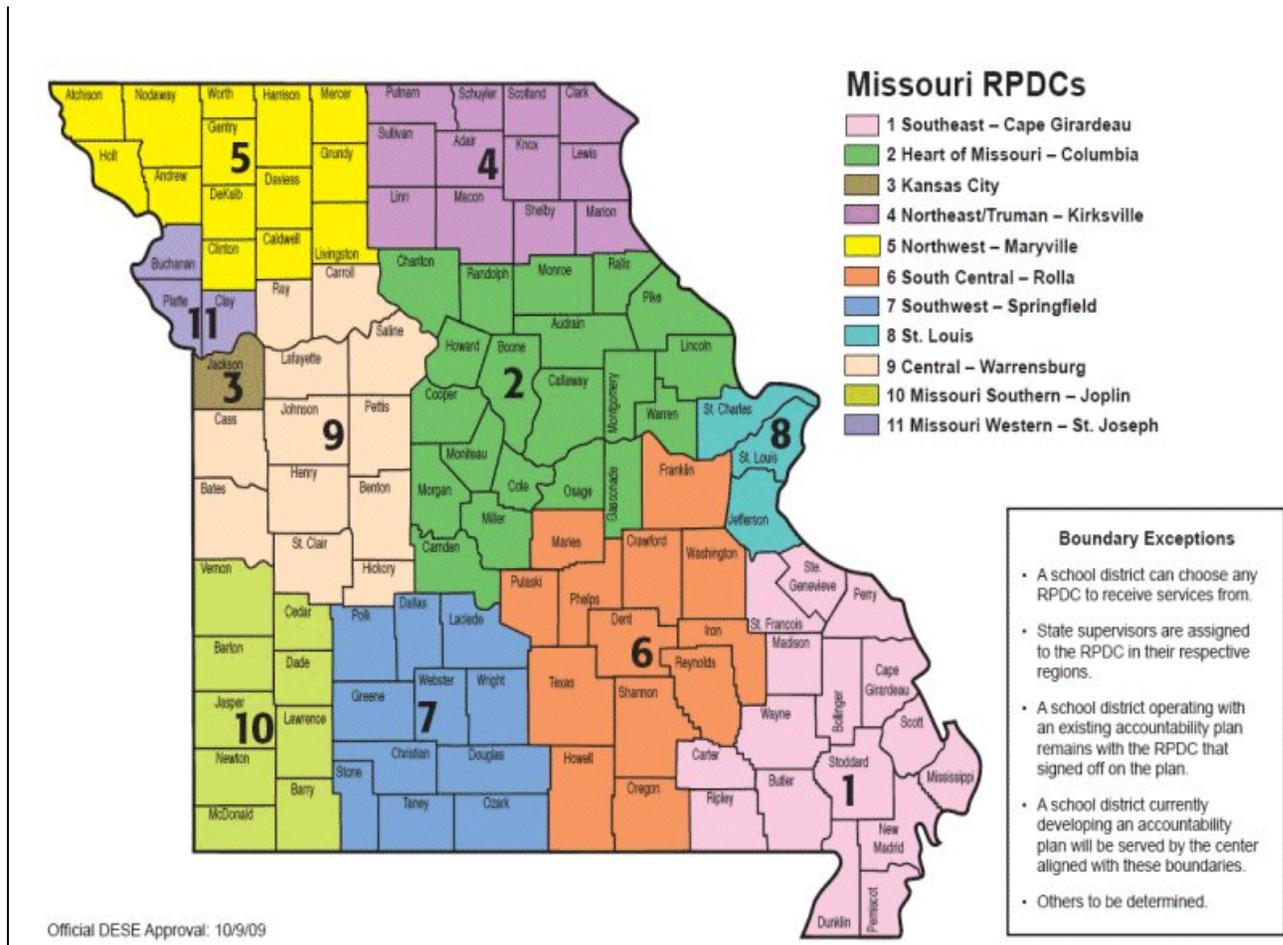


Figure 1. Missouri Regional Professional Development Centers Map

Materials

Panelists evaluated the alignment of the MAP items with the Missouri Grade-Level Expectations (GLEs). This section describes the GLEs reviewed, test form structure, and ratings forms and instructions used by panelists.

Test Forms. Panelists evaluated a single MAP 2010 test form and a 2011 test form per grade. Table 2.2 lists the characteristics of these test forms per grade-level test. This report does not include any examples of items or references to specific item content due to test security.

Table 2.2 Characteristics of 2010 and 2011 MAP Test Forms Reviewed

Content Area per Grade Level	2010 Test Forms			2011 Test Forms		
	Total Items per Form	Number of Selected Response Items	Number of Constructed Response/Performance Items	Total Items per Form	Number of Selected Response Items	Number of Constructed Response/Performance Items
Communication						
Arts						
3	57	52	5	56	51	5
4	56	52	4	55	51	4
5	55	51	4	55	51	4
6	56	52	4	55	51	4
7	61	56	5	61	56	5
8	60	56	4	60	56	4
Mathematics						
3	61	57	4	61	57	4
4	68	63	5	68	63	5
5	68	64	4	68	64	4
6	63	59	4	63	59	4
7	62	58	4	62	58	4
8	64	59	5	64	59	5
Science						
5	63	42	21	64	42	22
8	65	43	22	64	43	21

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

Procedures

HumRRO conducted this alignment review at the Assessment Resource Center (ARC) at the University of Missouri, Columbia, on October 6 through 8, 2009. While panels were convened in facilities procured through DESE, HumRRO directed the actual reviews independently of DESE. HumRRO provided workshop facilitators and group leaders for each small group content area panelists/reviewers. Group leaders were experienced in using Webb alignment process, use of the Webb Alignment Tool (WAT) and leading groups in doing alignment workshops, and were content area specialists. Prior to the workshop, facilitators and group leaders met to review procedures and materials.

The workshops began with introductions of staff and observers. Next, panelists read and signed affidavits of nondisclosure for the secure materials they would review during the workshop. HumRRO 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 be performing.

Following the general introduction, panelists began working within their content groups. The MAP reviewers were split into groups by content area and grade span. All groups contained seven reviewers. One HumRRO staff member helped facilitate each group.

Within their small groups, designated groups 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. HumRRO staff 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 two terminals and monitors. One monitor allowed for viewing of the PDF version of test items, while panelists used the second monitor to make ratings in the WAT.

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

Reviewers then received more specific instructions on rating items. For training, group leaders led panelists in evaluating and discussing sample items. After completing sample items, panelists rated each 2010 and 2011 test form item in the WAT. Panelists assigned a primary GLE to an item based on a judgment that an item clearly measured this content; however, reviewers could assign up to two additional GLEs if the item seemed to assess another standard equally to the primary standard. Panelists completed item ratings individually; however, group leaders led panelists through an adjudication process after all items on a form were completed to discuss any highly discrepant ratings. During the adjudication process, panelists were not required to come to 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 completed three types of surveys: (a) alignment summary comments (HumRRO), (b) feedback survey on alignment training and process (HumRRO), and (c) hotel accommodations survey (DESE).

Chapter 3 Results: Communication Arts

In this chapter, we report the results of the alignment review for Communication Arts. These analyses include inter-rater agreement and summary results on the four Webb alignment indicators. At the end of this chapter, we highlight and discuss key outcomes. Appendix C presents full, detailed statistical results on the Webb indicators, as well as tables with item-level results, items per GLE, and comments from reviewers.

Inter-rater Agreement Results

In this section, we report on two types of agreement analyses on panelists' ratings. Panelists rated the alignment of each item on two major dimensions: depth-of-knowledge and content match. The depth-of-knowledge rating required panelists to rank items using a scale, while the content rating involved a categorical judgment on the GLEs assessed by items. In either case, it is important to determine the extent to which panelists tended to provide exactly the same ratings on items. We applied a measure of absolute agreement to both types of panelists' ratings (Shavelson & Webb, N. M., 2005; Tinsley & Weiss, 1975).

For item DOK ratings, we applied the ICC (A, k) statistic, which refers to the intraclass correlation (ICC) coefficient used to measure the *absolute agreement* between panelists on scale ratings for items (see Brennan, 2001; Kane & Brennan, 1977; Putka & Sackett, in press). This statistic indicates the amount of agreement by producing a statistic between 0 and 1 (similar to a correlation coefficient). An ICC (A, 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

Table 3.1 presents inter-rater agreement outcomes for item DOK ratings (ICC). These results are listed separately for the 2010 and 2011 Communication Arts test forms per grade level.

Table 3.1 Interclass Correlation Coefficients on DOK Ratings for Communication Arts

Grade	ICC Results on DOK Ratings for 2010 Test Form	ICC Results on DOK Ratings for 2011 Test Form
3	0.93	0.91
4	0.93	0.91
5	0.95	0.94
6	0.92	0.88
7	0.94	0.90
8	0.91	0.93

The ICC (A, k) results in Table 3.1 indicate the reviewers applied the same DOK ratings to the same items frequently. All ICCs indicate 'Good' agreement between reviewers.

When evaluating agreement between categorical ratings such as GLE 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, we applied a measure developed by Norman Webb, which basically is an estimate of percent agreement between reviewers⁹. This analysis involves a pair-wise 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 pair-wise 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

Table 3.2 includes content match results at two levels of agreement. Columns 2 and 4 present exact agreement results, meaning agreement between reviewers at the Strand, Substrand, and GLE level. Columns 3 and 5 display results for partial agreement, meaning an assessment of agreement between reviewers at the Strand level only.

⁹ 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.

Table 3.2. Pair-Wise Comparisons on Content Agreement Between Communication Arts Reviewers

Grade	Pair-wise Comparisons on 2010 Test Forms		Pair-wise Comparisons on 2011 Test Forms	
	Exact Content Match (Strand, Big Idea, GLE)	Partial Content Match (Big Idea only)	Exact Content Match (Strand, Big Idea, GLE)	Partial Content Match (Big Idea only)
3	0.71	0.94	0.66	0.91
4	0.73	0.96	0.67	0.93
5	0.90	0.99	0.83	0.94
6	0.59	0.94	0.51	0.90
7	0.49	0.92	0.48	0.91
8	0.56	0.91	0.50	0.93

These results on pair-wise comparisons indicate that reviewers showed variable agreement on GLEs matched to items, particularly for exact matches on content strand, substrand, and GLEs. Such outcomes are not uncommon on measures of exact content match agreement between reviewers. However, a larger proportion of pair-wise comparisons than expected indicate ‘Weak’ agreement among reviewers, especially for the Grade 6, 7, and 8 panel in Communication Arts. These outcomes may have occurred for several reasons: inconsistency between reviewers in the application of GLEs to items; sufficient overlap in content for some GLEs to make discrete ratings more challenging; or, ambiguity in items on the target content assessed. In addition, the Grade 6-8 panel experienced several technical difficulties with the WAT, which required some adjustment to their process. Each of these factors can influence levels of agreement. The fact that these same reviewers showed higher agreement levels on item DOK ratings suggests that the decision criteria used to match items to GLEs may not have been a primary factor. Furthermore, reviewers made only a few notations about problematic items on each test form.

Webb Alignment Results

In this section, we review the general outcomes of item analyses on the four Webb alignment indicators. Detailed numeric results are found in Appendix A.

All of Webb’s measures begin with calculations for each reviewer and build up 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). However, 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 at these three sublevels per strand.

Categorical Concurrence

Categorical concurrence describes the extent to which the MAP items cover the content strands in the Missouri Grade-Level Expectations. Webb recommends a minimum of six test questions to adequately assess each content area (in this case, Big Idea). This criterion serves as a guideline for reasonable content coverage. Tables 3.3 and 3.4 summarize the MAP alignment results on categorical concurrence for the 2010 and 2011 test forms reviewed for Reading and Writing.

Table 3.3 Summary of Categorical Concurrence Results for Communication Arts-Reading and Writing per Big Idea, 2010 Test Form

Reading: Mean Number of Items per Big Idea for 2010 Test Form							
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms/Types	Strands with at Least Six Items
3	9.43	16.57	15.57	1.14	14.29	0.00	4 of 6
4	6.86	15.71	20.00	0.00	11.86	2.75	4 of 6
5	8.14	18.00	15.57	0.00	13.00	1.00	4 of 6
6	26.00	10.00	6.00	1.00	12.43	0.00	4 of 6
7	21.71	9.43	10.86	1.00	17.86	1.00	4 of 6
8	21.00	9.17	11.50	0.00	15.83	1.50	4 of 6

Note: Yellow shading indicates areas that did not meet the criterion.

Table 3.4 Summary of Categorical Concurrence Results for Communication Arts-Reading and Writing per Big Idea, 2011 Test Form

Reading: Mean Number of Items per Big Idea for 2011 Test Form							
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms/Types	Strands with at Least Six Items
3	9.14	21.57	9.57	1.67	14.29	1.25	4 of 6
4	8.86	11.14	19.86	1.00	11.43	3.17	4 of 6
5	8.43	13.14	19.29	0.00	13.00	1.00	4 of 6
6	26.43	8.86	5.00	1.00	14.29	1.00	3 of 6
7	23.14	10.00	9.86	1.00	17.14	0.00	4 of 6
8	22.29	7.14	11.43	1.00	17.29	2.00	4 of 6

Note: Yellow shading indicates areas that did not meet the criterion.

These results indicate that the 2010 and 2011 test forms include a sufficient number of items to cover the content adequately (minimum of 6 items per content area) for at least half of the Reading and Writing Big Ideas across grade levels. However, two Big Ideas (primarily Writing Process and Writing Forms/Types) did not receive adequate item coverage. Assessment of the writing process and modes often involves student demonstration of content knowledge through writing prompts or other constructed response items, of which there are few on most statewide assessments. If this explanation reflects the design of the Missouri MAP for Writing assessment, DESE could consider explicitly noting the scope of content eligible for constructed response

items in the Grade-Level Expectations or Test Specifications for more transparent documentation. Some states simply list possible content and note that specific content coverage on the assessment can vary per test cycle.

Depth-of-Knowledge Consistency

Analyses of depth of knowledge (DOK) measure the type of cognitive processing required of students by content standards. The DOK requirements implied by the GLEs should be matched by assessment items. To confirm this match, we asked panelists to rate the GLEs and the communication arts items separately. 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.

Tables 3.5 and 3.6 summarize the depth-of-knowledge consistency results at each grade level of the MAP for Reading and Writing. Because reviewers evaluated depth of knowledge at the most specific level of the standards document (GLEs), the table refers to consistency between the items and the GLEs to which they were matched. Results are summarized in terms of the percentage of items with cognitive complexity ratings at or above (more complex than) the rating for the corresponding GLE per Big Idea. Webb's suggested criterion for this alignment indicator is that at least 50% of the items should have complexity ratings at or above the level of the corresponding GLEs.

Table 3.5 Summary of Depth-of-Knowledge Results, Communication Arts, 2010 Test Form

Percent of 2010 Items with DOK At and Above the Level of the GLEs per Strand							
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms, Types	No. of Strands Assessed Adequately
3	38	78	62	93	79	0	4 of 6
4	74	78	86	0	89	0	4 of 6
5	77	90	18	0	92	100	4 of 6
6	22	71	41	0	76	0	2 of 6
7	28	48	44	0	73	67	2 of 6
8	24	60	36	0	74	50	3 of 6

Note: Yellow shading indicates areas that did not meet the criterion.

Table 3.6 Summary of Depth-of-Knowledge Results, Communication Arts, 2011 Test Form

Percent of 2011 Items with DOK At and Above the Level of the GLEs per Strand							
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms, Types	No. of Strands Assessed Adequately
3	36	77	71	67	82	100	5 of 6
4	71	72	88	0	84	25	4 of 6
5	75	89	19	0	91	75	4 of 6
6	20	63	50	0	74	0	3 of 6
7	25	46	38	0	76	0	1 of 6
8	25	54	44	0	66	67	3 of 6

Note: Yellow shading indicates areas that did not meet the criterion.

Reviewers rated the consistency between items and GLEs on DOK as rather low for most grade levels. In other words, many items assess students below the expected level of processing found in the GLEs. Assessment of the Writing Process clearly demonstrated the least amount of consistency with the GLEs – of the items assessing this content, only one grade-level test form each met the cognitive expectations of the corresponding GLEs. Other grade-level test forms showed varying degrees of consistency with the GLEs under the remaining Big Ideas.

Range-of-Knowledge Correspondence

The range-of-knowledge measure examines breadth of content coverage in greater detail. In addition to evaluating which content strands (or Big Ideas in this case) are assessed, we should consider how many of the GLEs within a strand are represented by items. Webb's minimum level of acceptability for range-of-knowledge correspondence is that at least 50% of GLEs per strand link with one or more items to ensure adequate breadth of content coverage *within* strands.

Table 3.7 lists the number of strands, Big Ideas, and GLEs found in the Missouri Grade-Level Expectations compared with the number of items per test form. This table only includes GLEs assessed on the MAP; additional locally assessed standards are not included in these counts.

Table 3.7. Number of Content Strands and GLEs Eligible for Assessment on MAP Communication Arts 2010 and 2011 Test Forms

Grade Level Test	Number of Content Strands	Number of Big Ideas	Number of GLEs Available for Assessment	Total Items for 2010 Form	Total Items for 2011 Form
3	2	6	16	57	56
4	2	6	16	56	55
5	2	6	17	55	55
6	2	6	16	56	55
7	2	6	16	61	61
8	2	6	16	60	60

To determine how many of these GLEs were matched to items, we first computed the frequency of GLEs covered (per Big Idea) separately for each panelist. Next, we calculated the mean number of GLEs linked with items across panelists. Tables 3.8 and 3.9 summarize the range-of-knowledge results for each grade level of the MAP for Reading and Writing. At least 50% of GLEs per strand should be assessed by one or more items for adequate coverage.

Table 3.8. Summary of Range-of-Knowledge Results, Communication Arts-Reading and Writing, 2010 Test Form

Reading: Mean Percent of GLEs per Big Idea for 2010 Test Form							
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms/Types	No. of Big Ideas Assessed Adequately
3	100	76	67	100	57	0	5 of 6
4	43	100	100	0	69	100	4 of 6
5	100	100	86	0	60	100	5 of 6
6	71	95	67	100	66	0	5 of 6
7	81	76	86	100	60	100	6 of 6
8	78	83	72	0	60	100	5 of 6

Note: Yellow shading indicates areas that did not meet the criterion.

Table 3.9. Summary of Range-of-Knowledge Results, Communication Arts-Reading and Writing, 2011 Test Form

Reading: Mean Percent of GLEs per Big Idea for 2011 Test Form							
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms & Types	No. of Big Ideas Assessed Adequately
3	100	71	48	100	54	100	5 of 6
4	52	95	95	<1	60	100	5 of 6
5	100	100	86	0	46	100	4 of 6
6	71	81	62	100	60	100	6 of 6
7	76	71	86	100	57	0	5 of 6
8	81	76	67	100	69	100	6 of 6

Note: Yellow shading indicates areas that did not meet the criterion.

These results indicate that the 2010 and 2011 test forms assess at least 50% of the GLEs per Big Idea with one or more items at many grade levels. Note that some Big Ideas (e.g., Writing- Text Development on 2010 Grade 3 form) barely met the minimum criterion for this alignment measure (M=57%). Furthermore, none of the GLEs under certain Big Ideas were assessed at all (Writing- Process on 2011 Grade 5 test form). Thus, these cases should be reviewed to determine whether item assignment could be redistributed or increased to improve alignment.

We provide a list of all GLEs matched to items by panelists in Appendix A.

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 GLE within each strand. The number of items should be distributed rather evenly between the GLEs 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 strand¹⁰. According to Webb, the minimum acceptable index for a single strand is 0.70 (on a scale of 0 to 1 with a 1 representing perfect balance). An index of 0.70 or higher suggests that items broadly assess the GLEs matched to items by reviewers instead of clustering around one or two GLEs¹¹.

¹⁰ 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>.

¹¹ 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 instead of the full pool of standards available for assessment.

One point should be noted regarding the balance index when interpreting the results. Only those GLEs actually matched to items by the panelists are included in calculations of the balance index. A given strand may include more GLEs than are actually linked to items by panelists. For example, if a particular strand includes eight GLEs in the state content standards document but panelists found items matching to just three GLEs, only these three GLEs 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.

Tables 3.10 and 3.11 summarize the results on balance-of-knowledge representation for each grade-level test form. An index of 0.70 or higher indicates adequate distribution of items among assessed GLEs.

Table 3.10. Summary of Balance-of-Knowledge Results, Communication Arts-Reading and Writing, 2010 Test Form

Reading: Mean Balance Index per Big Idea for 2010 Test Form							Big Ideas with Adequate Balance
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms/Types	
3	0.77	0.59	0.61	1.00	0.61	0 ¹	2 of 6
4	0.99	0.62	0.54	0	0.70	1.00	3 of 6
5	0.72	0.50	0.47	0	0.59	1.00	2 of 6
6	0.49	0.69	0.64	1.00	0.64	0	1 of 6
7	0.66	0.78	0.67	1.00	0.69	1.00	3 of 6
8	0.73	0.72	0.56	0	0.68	1.00	3 of 6

¹ No reviewer used the GLEs under this Big Idea.

Note: Yellow shading indicates areas that did not meet the criterion.

Table 3.11. Summary of Balance-of-Knowledge Results, Communication Arts-Reading and Writing, 2011 Test Form

Reading: Mean Balance Index per Big Idea for 2011 Test Form							Big Ideas with Adequate Balance
Grade	Reading Process	Reading Fiction	Reading Nonfiction	Writing Process	Writing Text Dev	Writing Forms/Types	
3	0.79	0.57	0.63	1.00	0.63	1.00	3 of 6
4	0.75	0.63	0.58	1.00	0.76	1.00	4 of 6
5	0.71	0.55	0.49	0	0.71	1.00	3 of 6
6	0.55	0.69	0.73	1.00	0.75	1.00	4 of 6
7	0.65	0.78	0.66	1.00	0.69	0	2 of 6
8	0.74	0.68	0.57	1.00	0.70	1.00	4 of 6

Note: Yellow shading indicates areas that did not meet the criterion.

The outcomes concerning item distribution among GLEs per Big Idea were mixed. Items assessing Reading seemed to be clustered around one or two GLEs per Big Idea across multiple grade levels. However, note that each Big idea for Reading

included a maximum of 3 GLEs; thus, small differences in item assignment affected the balance substantially. In contrast, the two Big Ideas under Writing (Processes and Forms/Types) included only one GLE each. When content strands only include a single content expectation, the result is usually a balance index of 1.00. Clearly, this outcome can be misconstrued when, in fact, the GLE may be linked to a single item. Both circumstances point to a need for caution in interpreting the balance outcomes for Communication Arts, particularly in light of some of the results for range-of-knowledge correspondence suggesting that a number of GLEs were not assessed for Writing Process and Writing Forms/Types in particular.

Summary and Discussion of Results on Webb Alignment Indicators

The content alignment review of the MAP evaluated the 2010 and 2011 test forms compared to the Missouri Grade-Level Expectations. 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 the No Child Left Behind Act of 2001.

HumRRO calculated the alignment results for Communication Arts at the level of the Big Idea (as opposed to the Strand level of Reading and Writing) so as not to obscure where alignment strengths and weaknesses may lie. The overall alignment results on the MAP test forms for Communication Arts suggest that some grade-level test forms align to the Missouri Grade-Level Expectations rather well on Webb measures, while other grade-level forms may require review of items to improve alignment. The outcomes for Writing (i.e., demonstration of Writing Processes and Writing Forms/Types) suggest that the assessment narrowly covers these topics. In comparison, the test forms covered Reading content more broadly, although the distribution of items may warrant review to reduce uneven assessment. Neither Reading nor Writing items assessed the GLEs at the appropriate depth-of-knowledge level. This conclusion applies to the majority of grade-level test forms.

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 at all of the alignment indicators together.

Tables 3.12 presents the summary alignment outcomes on the MAP Reading and Writing 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 Big Ideas that met the minimum alignment criteria. Table highlighting also corresponds with the scale above: green = highly to fully aligned; yellow = partially aligned (two-thirds majority); orange = partially aligned (exactly 50%); and, red = weakly aligned (less than 50%). This summary table links to the bottom row of tables in Appendix A (tables A-1 through A-24). Thus, these summary judgments reflect a final evaluation of each grade assessment per Webb criteria *across* the strands.

Table 3.12. Summary Degree of Alignment Outcomes per Webb Criterion for MAP Grade Level Tests in Communication Arts – Reading and Writing

Grade	2010 Test Form				2011 Test Form			
	Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
	CC	DOK	ROK	BOK	CC	DOK	ROK	BOK
3	High	Partial	High	Weak	Partial	Partial	High	Partial
4	Partial	Partial	Partial	Partial	Partial	Weak	Full	Partial
5	Partial	Partial	High	Partial	Partial	Partial	Partial	Partial
6	Partial	Weak	High	Weak	Partial	Partial	Full	Partial
7	Partial	Weak	Full	Partial	Partial	Partial	High	Weak
8	Partial	Partial	High	Partial	Partial	High	Full	Partial

Note: CC = Categorical Concurrence; DOK = Depth-of-knowledge Consistency; ROK = Range-of-knowledge Correspondence; BOK = Balance-of-knowledge Representation

Across the majority of grades, the 2010 and 2011 test forms appear to assess a good range of GLEs (range-of-knowledge correspondence). These positive outcomes (green highlighting) mostly apply to assessment of Big Ideas for Reading, as noted previously in Tables 3.8 and 3.9. Most alignment outcomes on remaining Webb criteria qualify as at least ‘partially aligned’ (shown in yellow highlighting), which in this case accounts for 4 of 6 Big Ideas. For example, the assessment exceeded the minimum criteria for categorical concurrence for all Big Ideas under Reading and one under Writing. The Writing Process and Writing Forms/Types Big Ideas received less emphasis overall in comparison.

Other outcomes reveal more serious alignment issues. For example, although a number of GLEs linked to items, it is still the case that the test forms disproportionately emphasized specific GLEs within strands/Big Ideas (balance-of-knowledge). In addition, many items did not assess corresponding GLEs at the appropriate level of cognitive processing expected in the Missouri Grade-Level Expectations, resulting in conclusions of lower alignment on depth-of-knowledge consistency. For these reasons, the test forms qualify as either ‘partially aligned’ at a lower level (item distribution among GLEs

for only 50% of Big Ideas was appropriate) or even weakly aligned (appropriate item distribution for between 1 and 3 Big Ideas) on these two Webb criteria. Orange highlighting was used to distinguish those grade test forms for which only half of Big Ideas met the minimum decision criteria.

Suggestions for improving the alignment between the Communication Arts-Reading and Writing assessments and Missouri Grade-Level Expectations are discussed in Chapter 6 Summary and Recommendations.

Chapter 4 Results: Mathematics

In this chapter, we report the results of the alignment review for Mathematics. These analyses include inter-rater agreement and summary results on the four Webb alignment indicators. At the end of this chapter, we highlight and discuss key outcomes. Appendix C presents full, detailed statistical results on the Webb indicators, as well as tables with item-level results, items per GLE, and comments from reviewers.

Inter-rater Agreement Results

In this section, we report on two types of agreement analyses based on panelists' ratings: (a) depth-of-knowledge and (b) content match. The depth-of-knowledge rating required panelists to rank items using a scale, while the content rating involved a categorical judgment on the GLEs assessed by items. In either case, it is important to determine the extent to which panelists tended to provide exactly the same ratings on items. We applied a measure of absolute agreement to both types of panelists' ratings (Shavelson & Webb, N. M., 2005; Tinsley & Weiss, 1975).

For item DOK ratings, we applied the ICC (A, k) statistic, which refers to the intraclass correlation (ICC) coefficient used to measure the *absolute agreement* between panelists on scale ratings for items (see Brennan, 2001; Kane & Brennan, 1977; Putka & Sackett, in press). This statistic indicates the amount of agreement by producing a statistic between 0 and 1 (similar to a correlation coefficient). An ICC (A, 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

Table 4.1 presents inter-rater agreement outcomes for item DOK ratings (ICC). These results are listed separately for the 2010 and 2011 Math test forms per grade level.

Table 4.1 Interclass Correlation Coefficients on DOK Ratings for Math

Grade	ICC Results on DOK Ratings for 2010 Test Form	ICC Results on DOK Ratings for 2011 Test Form
3	0.96	0.95
4	0.97	0.96
5	0.98	0.95
6	0.89	0.95
7	0.91	0.91
8	0.93	0.96

The ICC (A, k) results in Table 3.1 indicate the reviewers applied the same DOK ratings to the same items frequently. All ICCs indicate 'Good' agreement between reviewers.

When evaluating agreement between categorical ratings such as GLE 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, we applied a measure developed by Norman Webb, which basically is an estimate of percent agreement between reviewers¹². This analysis involves a pair-wise 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 pair-wise 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

Table 4.2 includes content match results at two levels of agreement. Columns 2 and 4 present exact agreement results, meaning agreement between reviewers at the Strand, Substrand, and GLE level. Columns 3 and 5 display results for partial agreement, meaning an assessment of agreement between reviewers at the Strand level only.

¹² 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.

Table 4.2. Pair-Wise Comparisons on Content Agreement for Math Reviewers

Grade	Pair-wise Comparisons on 2010 Test Forms		Pair-wise Comparisons on 2011 Test Forms	
	Exact Content Match (Strand, Substrand, GLE)	Partial Content Match (Strand only)	Exact Content Match (Strand, Substrand, GLE)	Partial Content Match (Strand only)
3	0.67	0.82	0.71	0.82
4	0.59	0.78	0.63	0.78
5	0.66	0.90	0.65	0.88
6	0.90	0.95	0.89	0.98
7	0.91	0.94	0.90	0.98
8	0.89	0.95	0.91	0.97

These results on pair-wise comparisons for content match indicate that reviewers showed ‘Adequate’ to ‘Good’ agreement on GLEs matched to items. Exact agreement between reviewers on content assessed by items was ‘Weak’ on the 2010 Grade 4 assessment. The reasons for the lower agreement on ratings of this test form are not clear.

Webb Alignment Results

In this section, we summarize outcomes of item analyses on the four Webb alignment indicators. Detailed numeric results are found in Appendix B.

All of Webb’s measures begin with calculations for each reviewer and build up 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 are presented at the strand level.

Categorical Concurrence

Categorical concurrence describes the extent to which the MAP items cover the content strands in the Missouri Grade-Level Expectations. Webb recommends a minimum of six test questions to adequately assess each content strand. This criterion serves as a guideline for reasonable content coverage. Tables 4.3 and 4.4 summarize the MAP alignment results on categorical concurrence for the 2010 and 2011 test forms reviewed.

Table 4.3 Summary of Categorical Concurrence Results, Math, 2010 Test Form

Mean Number of Items per Strand for 2010 Test Form						
Grade	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	Strands with at Least Six Items
3	15.71	11.29	10.86	9.43	6.57	5 of 5
4	18.43	19.14	10.71	12.86	7.71	5 of 5
5	15.57	14.29	10.57	8.86	9.14	5 of 5
6	18.29	12.86	8.57	8.57	13.43	5 of 5
7	15.71	19.43	8.29	6.86	9.57	5 of 5
8	13.14	20.43	15.57	5.86	13.00	4 of 5

Note: Yellow shading indicates areas that did not meet the criterion.

Table 4.4 Summary of Categorical Concurrence Results, Math, 2011 Test Form

Mean Number of Items per Strand for 2011 Test Form						
Grade	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	Strands with at Least Six Items
3	17.33	11.67	11.33	10.33	8.67	5 of 5
4	18.57	17.71	10.14	13.14	7.71	5 of 5
5	15.57	14.29	10.57	8.86	9.14	5 of 5
6	17.00	12.86	8.00	9.00	14.43	5 of 5
7	15.71	18.00	9.00	11.14	10.29	5 of 5
8	11.71	20.29	16.43	6.29	13.00	5 of 5

These results ($M > 6$ in each case) indicate that the MAP test forms adequately cover the breadth of the Math content strands that students are expected to know across grade levels, except for the Measurement strand for the 2010 test form.

Depth-of-Knowledge Consistency

Analyses of depth-of-knowledge (DOK) measure the type of cognitive processing required of students by content standards. The DOK requirements implied by the GLEs should be matched by assessment items. To confirm this match, panelists were asked to rate the GLEs and the mathematics items separately. 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 a rating scale (from Webb, 2005) with four levels of cognitive complexity. Appendix E includes further information and examples of the DOK levels.

- 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.

Tables 4.5 and 4.6 summarize the depth-of-knowledge consistency results for each grade level of the MAP. Because reviewers evaluated depth-of-knowledge at the most specific level of the standards document (GLEs), the table refers to consistency between the items and the GLEs to which they were matched. Results are summarized in terms of the percentage of items with cognitive complexity ratings at or above (more complex than) the rating for the corresponding GLE. Webb's suggested criterion for this alignment indicator is that at least 50% of the items should have complexity ratings at or above the level of the corresponding GLE.

Table 4.5. Summary of Depth-of-Knowledge Results, Math, 2010 Test Form

Percent of 2010 Items with DOK At and Above the Level of the GLEs per Strand						
Grade	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	Number of Strands Assessed Adequately
3	70	87	77	84	89	5 of 5
4	92	79	84	81	81	5 of 5
5	79	83	70	100	54	5 of 5
6	72	72	67	66	39	4 of 5
7	64	51	92	74	64	5 of 5
8	71	77	76	54	46	4 of 5

Note: Yellow shading indicates areas that did not meet the criterion.

Table 4.6 Summary of Depth-of-Knowledge Results, Math, 2011 Test Form

Grade	Percent of 2011 Items with DOK At or Above the Level of the GLEs per Strand					Number of Strands Assessed Adequately
	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	
3	75	88	79	84	77	5 of 5
4	86	74	78	79	90	5 of 5
5	77	88	73	100	68	5 of 5
6	65	74	81	55	47	4 of 5
7	91	68	70	91	63	5 of 5
8	74	94	75	49	30	3 of 5

Note: Yellow shading indicates areas that did not meet the criterion.

Panelists' ratings on depth-of-knowledge consistency for the Math tests forms suggest that many MAP items assess students at the level expected in the Missouri Grade-Level Expectations. The exceptions were Grades 6 and 8 for Data and Probability and Grade 8 for Measurement, where less than 50% of items assessed students at the same cognitive levels expected in the Data and Probability GLEs. This outcome indicates that reviewers rated multiple items as below the corresponding GLEs for these strands.

Range of Knowledge

The range-of-knowledge measure examines breadth of knowledge in greater detail. In addition to evaluating which content strands are assessed, we should consider how many of the GLEs within a strand are represented by items. The GLEs 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 GLEs per strand link with items to ensure adequate breadth of content coverage *within* strands.

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

Table 4.7. Number of Content Strands and GLEs Eligible for Assessment on Math 2010 and 2011 MAP Test Forms

Grade Level Test	Number of Content Strands	Number of GLEs Available for Assessment per Grade	Total Items for 2010 Form	Total Items for 2011 Form
3	5	17	61	61
4	5	20	68	68
5	5	18	68	68
6	5	23	63	63
7	5	23	62	62
8	5	25	64	64

To determine how many of these GLEs were matched to items, we first computed the frequency of GLEs covered (per strand) separately for each panelist. Next, we calculated the mean number of GLEs linked with items across panelists. Tables 4.8 and 4.9 summarize the range-of-knowledge results for each grade level of the MAP per content strand. At least 50% of GLEs per strand should be assessed by one or more items for adequate coverage.

Table 4.8. Summary of Range-of-Knowledge Results, Math, 2010 Test Form

Percent of GLEs per Strand with Assessed by At Least One Item on 2010 Test Form						
Grade	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	Number of Strands Assessed Adequately
3	91	90	96	96	100	5 of 5
4	90	93	100	100	100	5 of 5
5	92	86	100	69	100	5 of 5
6	84	100	100	100	100	5 of 5
7	74	100	82	100	68	5 of 5
8	96	88	88	67	100	5 of 5

Table 4.9. Summary of Range-of-Knowledge Results, Math, 2011 Test Form

Percent of GLEs per Strand with Assessed by At Least One Item on 2011 Test Form						
Grade	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	Number of Strands Assessed Adequately
3	95	96	100	100	100	5 of 5
4	90	93	100	100	93	5 of 5
5	78	82	100	95	100	5 of 5
6	82	100	100	100	100	5 of 5
7	78	100	82	100	76	5 of 5
8	96	84	94	67	100	5 of 5

Results for both test forms per grade indicate that items adequately covered a range of GLEs for each strand. Thus, reviewers matched most to all of the GLEs to items. Note that, while the outcome for GLEs in the Measurement strand appears to suggest much lower coverage by items relative to the other strands, this strand includes only three GLEs total.

As a final comparison of items GLEs, Table 4.10 lists the mean total number of GLEs matched to items (per grade level and across strands) relative to the actual total GLEs per grade level. These results further confirm that the test forms assess the majority of Math GLEs.

Table 4.10. Comparison of GLEs Matched to Items with GLEs Available for Assessment on Math 2010 and 2011 MAP Test Forms

Grade	Number of GLEs Available for Assessment	2010 Forms		2011 Forms	
		Mean Number of GLEs Matched to Items by Panelists	Mean Percentage of GLEs Matched to Items by Panelists	Mean Number of GLEs Matched to Items by Panelists	Mean Percentage of GLEs Matched to Items by Panelists
3	17	16.56	97	17.49	> 100 ^a
4	20	19.99	100	19.85	99
5	18	17.00	94	16.86	94
6	23	22.86	99	23.00	100
7	23	19.71	86	24.71	> 100
8	25	23.00	92	22.57	90

^a Totals of greater than 100% indicate some items were matched to more than one GLE.

We provide a list of all GLEs matched to items by panelists in Appendix B if DESE or CTB wishes to review items matched to each GLE.

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 GLE within each strand. The number of items should be distributed rather evenly between the GLEs 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 strand¹³. According to Webb, the minimum acceptable index for a single strand is 0.70 (on a scale of 0 to 1 with a 1 representing perfect balance). An index of 0.70 or higher

¹³ 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>.

suggests that items broadly assess the GLEs matched to items by reviewers instead of clustering around one or two GLEs¹⁴.

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

Tables 4.11 and 4.12 summarize the results on balance-of-knowledge representation for each grade-level test form. An index of 0.70 or higher indicates adequate distribution of items among assessed GLEs

Table 4.11. Summary of Balance-of-Knowledge Results, Math, 2010 Test Form

Balance Index per Strand for 2010 Test Form						
Grade	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	Strands with Adequate Balance
3	0.81	0.79	0.86	0.76	1.00	5 of 5
4	0.82	0.75	0.81	0.75	0.78	5 of 5
5	0.78	0.82	0.87	0.85	0.92	5 of 5
6	0.60	0.74	0.77	0.87	0.80	4 of 5
7	0.67	0.72	0.79	0.82	0.68	3 of 5
8	0.75	0.73	0.77	0.80	0.71	5 of 5

Note: Yellow shading indicates areas that did not meet the criterion.

Table 4.12. Summary of Balance-of-Knowledge Results, Math, 2011 Test Form

Balance Index per Strand for 2011 Test Form						
Grade	Numbers and Operations	Algebraic Relationships	Geometric and Spatial Relationships	Measurement	Data and Probability	Strands with Adequate Balance
3	0.82	0.73	0.82	0.81	1.00	5 of 5
4	0.75	0.67	0.83	0.79	0.87	4 of 5
5	0.86	0.80	0.83	0.75	0.88	5 of 5
6	0.65	0.77	0.90	0.78	0.78	4 of 5
7	0.66	0.70	0.81	0.71	0.64	3 of 5
8	0.75	0.65	0.73	0.86	0.64	3 of 5

Note: Yellow shading indicates areas that did not meet the criterion.

¹⁴ 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 instead of the full pool of standards available for assessment.

The results for the Math 2010 and 2011 test forms indicate that items are distributed across GLEs in a relatively even manner on most grade-level forms. The outcomes previously described in Table 4.10 regarding the number of GLEs matched to items seem to support this conclusion.

DESE and CTB may want to review item assignment relative to the Numbers and Operations strand and the Data and Probability strand for Grades 6 and 7. In addition, reviewer ratings of the 2011 test form in particular suggest that these operational items may cluster around a small number of GLEs for the Algebraic Relationships strand at Grades 4 and 8 as well. The Grade 6 outcomes for Numbers and Operations on Range and Balance provide one example. In reviewing the detailed grade-level results in Appendix B, Tables B-16 and B-17 show that panelists matched approximately 18 items on the Grade 6 test forms to five of the Numbers and Operations GLEs (2010 Mean Items = 18.29; Mean GLEs = 5.86). Of those five GLEs under Numbers and Operations, reviewers matched the greatest number of items to the GLE N.3.e.6, followed by GLE N.3.c.6. The remaining Numbers and Operations GLEs were matched to one item each. While the Missouri test blueprint for Grade 6 Math indicates that the Numbers and Operations strand should receive greater emphasis (which is common and reasonable in state content standards), it is not clear why these two particular GLEs within this strand are assessed disproportionately on the test forms. Further explanation by DESE may justify this emphasis.

Summary and Discussion of Results on Webb Alignment Indicators

The content alignment review of the MAP evaluated the 2010 and 2011 test forms compared to the Missouri Grade-Level Expectations. 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 the No Child Left Behind Act of 2001.

The overall alignment results for the MAP test forms for Math suggest that test items align to the GLEs at a high level. Results suggest that the Grades 7 and 8 test forms, however, may over-emphasize some GLEs. We present summary alignment judgments across strands per grade level based on the statistical outcomes.

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%) (green);
- Highly aligned – assessments align to the majority of strands (70%–90%) (green)
- Partially aligned – assessments align well to some strands (50%–69%) (yellow, greater than 50% to 69%; orange, 50%)

- Weakly aligned – assessments align few strands (below 50%) (red).

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 at all of the alignment indicators together.

Table 4.13 presents the summary alignment outcomes for the MAP 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 in Appendix B (Tables B-1 through B-24). Thus, these summary judgments reflect a final evaluation of each grade assessment per Webb criteria *across* the strands.

Table 4.13. Summary Alignment Outcomes per Webb Criterion for MAP Grade Level Tests in Math

Grade	2010 Test Form				2011 Test Form			
	Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
	CC	DOK	ROK	BOK	CC	DOK	ROK	BOK
3	Full	Full	Full	Full	Full	Full	Full	Full
4	Full	Full	Full	Full	Full	Full	Full	High
5	Full	High	Full	Full	Full	Full	Full	Full
6	Full	High	Full	High	Full	High	Full	High
7	Full	Full	Full	Partial	Full	Full	Full	Partial
8	Full	Full	Full	Full	Full	Partial	Full	Partial

Note: Yellow shading indicates areas that did not meet the criterion.

As shown in Table 4.13 with green highlighting, most alignment results point to good content alignment of the MAP to the Missouri Grade-Level Expectations. Reviewers’ ratings indicate that each grade-level test form includes a sufficient number of operational items to cover all content strands (categorical concurrence) as well as a range of GLEs within those strands (range-of-knowledge correspondence).

Results on the Grade 7 test form seem to suggest that a number of items cluster around one or two GLEs for the Numbers and Operations strand and the Data and Probability strand in particular (noted in yellow highlighting). While reviewers matched most GLEs within these strands to items, the balance-of-knowledge representation results suggest that the majority of items assess two GLEs. Grade 8 exhibits a similar pattern for the Algebraic Relationships strand and Data and Probability strand. In this case, items tend to cluster primarily around one GLE. Thus, the 2010 and 2011 test forms only partially align to the GLEs for these grades.

Recommendations and suggestions for improving alignment between the Math assessments and Missouri Grade-Level Expectations are discussed in Chapter 6 Summary and Recommendations.

Chapter 5 Results: Science

In this chapter, we report the results of the alignment review for Science. These analyses include inter-rater agreement and summary results on the four Webb alignment indicators. At the end of this chapter, we highlight and discuss key outcomes. Appendix C presents full, detailed statistical results on the Webb indicators, as well as tables with item-level results, items per GLE, and comments from reviewers.

Inter-rater Agreement Results

In this section, we report on two types of agreement analyses on panelists' ratings. Panelists rated the alignment of each item on two major dimensions: depth-of-knowledge and content match. The depth-of-knowledge rating required panelists to rank items using a scale, while the content rating involved a categorical judgment on the GLEs assessed by items. In either case, it is important to determine the extent to which panelists tended to provide exactly the same ratings on items. We applied a measure of absolute agreement to both types of panelists' ratings (Shavelson & Webb, N. M., 2005; Tinsley & Weiss, 1975).

For item DOK ratings, we applied the ICC (A, k) statistic, which refers to the intraclass correlation (ICC) coefficient used to measure the *absolute agreement* between panelists on scale ratings for items (see Brennan, 2001; Kane & Brennan, 1977; Putka & Sackett, in press). This statistic indicates the amount of agreement by producing a statistic between 0 and 1 (similar to a correlation coefficient). An ICC (A, 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

Table 5.1 presents inter-rater agreement outcomes for item DOK ratings (ICC). These results are listed separately for the 2010 and 2011 Science test forms per grade level.

Table 5.1 Interclass Correlation Coefficients on DOK Ratings for Science

Grade	ICC Results on DOK Ratings for 2010 Test Form	ICC Results on DOK Ratings for 2011 Test Form
5	0.93	0.91
8	0.92	0.85

The ICC (A, k) results in Table 5.1 indicate the reviewers applied the same DOK ratings to the same items frequently. All ICCs indicate 'Good' agreement between reviewers.

When evaluating agreement between categorical ratings such as GLE 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, we applied a measure developed by Norman Webb, which basically is an estimate of percent agreement between reviewers¹⁵. This analysis involves a pair-wise 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 pair-wise 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

Table 5.2 includes content match results at two levels of agreement. Columns 2 and 4 present exact agreement results, meaning agreement between reviewers at the Strand, Substrand, and GLE level. Columns 3 and 5 display results for partial agreement, meaning an assessment of agreement between reviewers at the Strand level only.

Table 5.2. Pair-Wise Comparisons on Content Agreement Between Reviewers for Science, 2010 and 2011 Test Forms

Grade	Pair-wise Comparisons on 2010 Test Forms		Pair-wise Comparisons on 2011 Test Forms	
	Exact Content Match (Strand, Substrand, GLE)	Partial Content Match (Strand only)	Exact Content Match (Strand, Substrand, GLE)	Partial Content Match (Strand only)
5	0.72	0.93	0.77	0.96
8	0.72	0.93	0.78	0.96

These results on pair-wise comparisons indicate that reviewers showed good agreement on GLEs matched to items even for exact matches across content strand,

¹⁵ 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.

substrand, and GLEs. While exact agreement to the GLE level is lower relative to the strand level, the outcomes in Table 5.2 indicate that reviewers frequently assigned the same GLEs to the same items.

Webb Alignment Results

In this section, we review the general outcomes of item analyses on the four Webb alignment indicators. Detailed numeric results are found in Appendix C.

All of Webb's measures begin with calculations for each reviewer and build up 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 are presented at the strand level.

Findings for Science require some additional explanation for the following reasons. In matching items to Science GLEs, panel members matched items against the following choices: all strands, concepts, and assessed GLEs for the grade span (Grade Span 3-5 and Grade Span 6-8). As a result, when the panel members then matched the items to the GLEs, they could match an item to any of the grade span's assessed GLEs. For Grade 5, panel members matched the 63 items from the 2010 form and 64 items on the 2011 form to the 149 rated strands, concepts, and assessed GLEs for Grade Span 3-5. If panel members were restricted to matching against only those GLEs assessed for Grade 5, they would have been matching the items to only 56 GLEs. Grade 8 procedures were the same, but for Grade Span 6-8. Approximately 16% of the panelists' ratings matched items to Grade 3 standards, 22% to Grade 4 standards, and 62% to Grade 5 standards for the Grade 5 assessment. Approximately 24% of panelists' ratings matched items to Grade 6 standards, 27% to Grade 7 standards, and 50% to Grade 8 standards on the Grade 8 assessment. All panelists ratings are provided in the Appendices Tables C-13 and C-14.

The decision to have panelists match items to the grade span rather than to a single grade most directly affects the results in range of knowledge. For range of knowledge, only 1 strand was found to be adequately assessed from the possible 32 Science strands across both forms and grade levels. However, matching the 63 and 64 items for Grade 5 forms 2010 and 2011, respectively, to 56 rather than 149 choices provides a far higher likelihood for matching at least 50% of the GLEs within each strand. The same logic holds for matching the 65 and 64 items for Grade 8 forms 2010 and 2011, respectively, to 82 rather than 219 choices.

Categorical concurrence should not be affected since the strands remain constant across all grades.

Depth of knowledge may have been affected, but not to a large extent. The impact would probably be that more items were at the same or higher DOK level as the standards since the items also were being matched to lower grades' GLEs rather than only for Grades 5 and 8 (assuming that there was a tendency for standards to have

lower DOK requirements at lower grades). This may have actually increased the number of strands that were determined to be adequately assessed.

Balance of knowledge also may be higher as a result of the items being dispersed over more choices. A quick scan of the data matches for Grades 5 and 8 found that the most frequently selected choices had only one or two items matched and very few choices with more than three items matched.

However, it is not possible to examine the data to match only with the Grade 5 and Grade 8 GLEs. Panelists matched items to standards from the grade span, and we cannot ascertain what standards they would have matched, or if they would have matched standards at all, if only a single grade’s standards had been presented. All analyses and results for Science in this report were based on the assumption that each assessment represented a three-grade span.

Categorical Concurrence

Categorical concurrence describes the extent to which the MAP items cover the content strands in the Missouri Grade-Level Expectations. Webb recommends a minimum of six test questions to adequately assess each content strand. This criterion serves as a guideline for reasonable content coverage. Tables 5.3 and 5.4 summarize the MAP alignment results on categorical concurrence for the 2010 and 2011 test forms reviewed¹⁶.

Table 5.3. Summary of Categorical Concurrence Results, Science, 2010 Test Form

Mean Number of Items per Strand for 2010 Test Form									
Grade	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	Strands with at Least Six Items
5	11.14	7.71	7.14	8.86	10.86	8.00	23.86	5.29	7 of 8
8	8.86	6.14	5.43	5.57	8.43	6.86	19.57	3.14	5 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

Table 5.4. Summary of Categorical Concurrence Results, Science, 2011 Test Form

Mean Number of Items per Strand for 2011 Test Form									
Grade	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	Strands with at Least Six Items
5	12.86	4.71	6.71	9.14	11.86	5.71	20.71	5.71	5 of 8
8	13.17	3.00	12.50	6.00	13.33	7.83	26.33	6.83	7 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

¹⁶ Since strands remained the same for across the grade spans, this criterion should not have been impacted by matching to all GLEs across the grade span.

These results indicate that most MAP test forms adequately cover the breadth of the Science content strands that students are expected to know across grade levels. However, several strands received less emphasis, as shown by means of less than six items. The number of items assigned by reviewers to the Science and Technology strand in particular roughly corresponds with the test blueprint (6 – 8 points). The remaining strands with means below six items fall well below the test blueprint targets, however. Thus, reviewers' judgments about which items correspond to the GLEs does not correspond with the goals of the test developer for these strands.

Depth-of-Knowledge Consistency

Analyses of depth of knowledge (DOK) measure the type of cognitive processing required of students by content standards. The DOK requirements implied by the GLEs should be matched by assessment items. To confirm this match, panelists were asked to rate the GLEs and the Science items separately. 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 a rating scale (adapted from Webb, 2005) with four levels of cognitive complexity. Further information and examples of the DOK levels are found in Appendix E.

- | | |
|---|--|
| <ul style="list-style-type: none"> • Level 1 Recognition • Level 2 Skills/Concepts • Level 3 Strategic Thinking • Level 4 Extended Thinking | <ul style="list-style-type: none"> - simple recall of information (i.e., facts, terms); sequencing; more automatic. - beyond habitual response; applying concepts; problem-solving. - requires basic reasoning, planning, or use of evidence; generating hypotheses. - complex reasoning; evaluation of multiple sources or independent pieces of evidence; often over an extended period of time. |
|---|--|

Tables 5.5 and 5.6 summarizes the depth-of-knowledge consistency results for each grade level of the MAP¹⁷. Because reviewers evaluated depth of knowledge at the most specific level of the standards document (GLEs), the table refers to consistency between the items and the GLEs to which they were matched. Results are summarized in terms of the percentage of items with cognitive complexity ratings at or above (more complex than) the rating for the corresponding GLE. Webb's suggested criterion for this alignment indicator is that at least 50% of the items should have complexity ratings at or above the level of the corresponding GLE.

¹⁷ The results for this criterion should not have been adversely impacted by examining items to all GLEs within each grade span.

Table 5.5. Summary of Depth-of-Knowledge Results, Science, 2010 Test Form

Percent of 2010 Items with DOK At or Above the Level of the GLEs per Strand									
Grade	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	Number of Strands Assessed Adequately
5	81	79	90	79	76	100	67	54	8 of 8
8	57	59	85	45	52	44	81	21	5 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

Table 5.6 Summary of Depth-of-Knowledge Results, Science, 2011 Test Form

Percent of 2011 Items with DOK At or Above the Level of the GLEs per Strand									
Grade	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	Number of Strands Assessed Adequately
5	78	100	97	71	76	79	39	54	7 of 8
8	38	83	67	42	44	30	86	32	3 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

The tables above indicate that the 2010 and 2011 test forms show mixed results on DOK consistency. The 2010 Grade 5 test form adequately assesses student knowledge at a comparable cognitive level as expected in the Missouri Grade-Level Expectations, although the outcome for Science and Technology suggests that the DOK level for just over half of items matches the DOK level of corresponding GLEs. The 2011 Grade 5 test form shows lower consistency with the Scientific Inquiry GLEs, and, again, just over half of items matched the Science and Technology GLEs on cognitive level assessed.

Results on DOK consistency for both of the Grade 8 test forms reveal that many items assess students below the cognitive level of the corresponding GLEs for several strands. For a number of items, the discrepancy between the item DOK and the DOK of the GLEs involves adjacent ratings. For example, reviewers assigned a DOK rating at level 1 (recall and recognition) to the item, while the corresponding GLE expects performance at DOK level 2 (demonstration of skills and concepts). Adjacent ratings (DOK 1 versus 2) reflect less critical discrepancies than if ratings deviated by two or more scale values (DOK of 1 versus 3). DESE may wish to review the items targeting those strands in Grade 8 for which DOK consistency results indicate low alignment to determine if items require modification to meet the GLEs in some cases.

Range of Knowledge

The range-of-knowledge measure examines breadth of knowledge in greater detail. In addition to evaluating which content strands are assessed, we should consider how many of the GLEs within a strand are represented by items. The GLEs 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 GLEs per strand link with items to ensure adequate breadth of content coverage *within* strands.

Table 5.7 lists the number of strands and GLEs found in the Missouri Grade-Level Expectations compared with the number of items per test form. This table only includes GLEs assessed on the MAP; additional locally assessed standards are not included in these counts.

Table 5.7. Number of Content Strands and GLEs Eligible for Assessment on MAP Science 2010 and 2011 Test Forms

Grade Level Test	Number of Content Strands	Number of GLEs Available for Assessment per Grade	Total Items for 2010 Form	Total Items for 2011 Form
5	8	149	63	64
8	8	219	65	64

Note: The number of GLEs listed in this table reflect the combined number of GLEs across the grade span. Thus, for Grade 5, these are the GLEs for Grades 3-5. For Grade 8, the GLEs are for Grades 6-8.

To determine how many of these GLEs were matched to items, we first computed the frequency of GLEs covered (per strand) separately for each panelist. Next, we calculated the mean number of GLEs linked with items across panelists. Tables 5.8 and 5.9 summarize the range-of-knowledge results for each grade level of the MAP per content strand¹⁸. At least 50% of GLEs per strand should be assessed by one or more items for adequate coverage.

Table 5.8. Summary of Range-of-Knowledge Results, Science, 2010 Test Form

Grade	Percent of GLEs per Strand Assessed by At Least One Item on 2010 Test Form								Number of Strands Assessed Adequately
	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	
5	23	27	35	34	36	24	43	33	0 of 8
8	15	29	15	30	21	20	44	39	0 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

¹⁸ These results are for grade-level items matched to all GLEs with the grade span (3-5 and 6-8). If the grade-level items are matched only to Grade 5 or Grade 8 GLEs, there would be a far greater likelihood of matching at least 50% of the GLEs within each strand.

Table 5.9. Summary of Range-of-Knowledge Results, Science, 2011 Test Form

Percent of GLEs per Strand Assessed by At Least One Item on 2011 Test Form									
Grade	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	Number of Strands Assessed Adequately
5	25	22	29	29	84	24	41	29	1 of 8
8	16	15	24	32	22	19	48	46	0 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

Results indicate that the 2010 and 2011 test forms for Grades 5 and 8 assess a small set of GLEs for every Science strand. The one exception is the Grade 5 items targeting the Earth System GLEs (M = 84% of GLEs matched to at least one item).

Several issues should be considered as explanation for the weak alignment outcomes on range-of-knowledge correspondence. First, a substantial amount of content is available for assessment, as demonstrated by the list of GLEs in Table 3.8. As the number of specific content expectations increase, the ability of the test to adequately cover these expectations decreases due to practical limits in test length. Clearly, a test form with a maximum of 64 points cannot reasonably assess all 149 GLEs, for example. Second, although Missouri does prioritize some GLEs for assessment (i.e., Scientific Inquiry receives at twice as many points compared to other strands) in the Test Specifications, DESE may wish to consider options for reducing the number of GLEs requiring assessment.

As further demonstration of the disproportionate number of GLEs to assessment items, Table 5.10 shows the mean total number of GLEs matched to items (per grade level and across strands) relative to the actual total GLEs per grade level.

Table 5.10. Comparison of GLEs Matched to Items with GLEs Available for Assessment for 2010 and 2011 MAP Science Test Forms

Grade	Number of GLEs Available for Assessment	2010 Forms		2011 Forms	
		Mean Number of GLEs Matched to Items by Panelists	Mean Percentage of GLEs Matched to Items by Panelists	Mean Number of GLEs Matched to Items by Panelists	Mean Percentage of GLEs Matched to Items by Panelists
5	149	47.00	32%	45.48	31%
8	219	48.70	22%	52.17	24%

We provide a list of all GLEs, including those matched to items by panelists, in Appendix C for further review.

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 GLE within each strand. The number of items should be distributed rather evenly between the GLEs 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 strand¹⁹. According to Webb, the minimum acceptable index for a single strand is 0.70 (on a scale of 0 to 1 with a1 representing perfect balance). An index of 0.70 or higher suggests that items broadly assess the GLEs matched to items by reviewers instead of clustering around one or two GLEs²⁰.

One point should be noted regarding the balance index when interpreting the results. Only those GLEs actually matched to items by the panelists are included in calculations of the balance index. A given strand may include more GLEs than are actually linked to items by panelists. For example, if a particular strand includes eight GLEs in the state content standards document but panelists found items matching to just three GLEs, only these three GLEs 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.

Tables 5.11 and 5.12 summarize the results on balance-of-knowledge representation for each grade-level test form²¹. An index of 0.70 or higher indicates adequate distribution of items among assessed GLEs

Table 5.11. Summary of Balance-of-Knowledge Results, Science, 2010 Test Form

Balance Index per Strand for 2010 Test Form									
Grade	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	Strands with Adequate Balance
5	0.84	0.85	0.77	0.87	0.80	0.86	0.72	0.91	8 of 8
8	0.85	0.96	0.86	0.91	0.82	0.78	0.67	0.86	7 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

¹⁹ 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>.

²⁰ 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 instead of the full pool of standards available for assessment.

²¹ Depth of knowledge results may be impacted by matching items to grade span GLEs rather than grade-specific GLEs.

Table 5.12. Summary of Balance-of-Knowledge Results, Science, 2011 Test Form

Grade	Balance Index per Strand for 2011 Test Form								Strands with Adequate Balance
	Matter and Energy	Force and Motion	Living Organisms	Ecology	Earth Systems	Universe	Scientific Inquiry	Science and Technology	
5	0.81	0.87	0.78	0.83	0.81	0.78	0.70	0.89	8 of 8
8	0.81	0.97	0.84	0.87	0.79	0.78	0.63	0.84	7 of 8

Note: Yellow shading indicates areas that did not meet the criterion.

The results suggest that the Grades 5 and 8 test forms rather evenly represent the GLEs across strands, except for Scientific Inquiry on the 2011 Grade 5 and Grade 8 forms. As noted previously, however, these outcomes must be interpreted within the context of the range-of-knowledge correspondence results, which indicated that a small number of GLEs matched to items overall. We can at least conclude that, of the GLEs actually assessed, item representation is good for seven of eight Science strands.

Summary and Discussion of Results on Webb Alignment Indicators

The content alignment review of the MAP evaluated the 2010 and 2011 Science test forms compared to the Missouri Grade-Level Expectations. 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 the No Child Left Behind Act of 2001.

HumRRO applied the Webb alignment method to conduct the review. The overall alignment results for the MAP were mixed. The 2010 Science test form for Grade 5 exhibits good alignment on three Webb indicators, and the other test forms for Grades 5 and 8 indicate adequate alignment on several Webb indicators each. We present summary alignment judgments across strands per grade level based on the statistical outcomes.

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 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 at all of the alignment indicators together.

Table 5.13 presents the summary alignment outcomes for the MAP based on the above scale²². 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 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.

The highlighting in Table 5.13 denotes areas of good (green), moderate (yellow), and weak (red) alignment.

Table 5.13. Summary Alignment Outcomes per Webb Criterion for MAP Science Tests

Grade	2010 Test Form				2011 Test Form			
	Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
	CC	DOK	ROK	BOK	CC	DOK	ROK	BOK
5	High	Full	Weak	Full	Partial	High	Weak	Full
8	Partial	Partial	Weak	High	High	Weak	Weak	High

Note: CC = Categorical Concurrence; DOK = Depth-of-knowledge Consistency; ROK = Range-of-knowledge Correspondence; BOK = Balance-of-knowledge Representation

Based on the alignment conclusions in Table 5.13, the 2010 and 2011 test forms for Grades 5 and 8 appear to warrant further review, although the Grade 8 forms appear to show more comprehensive alignment issues to the Science GLEs overall. Inconsistent assessment of the Scientific Inquiry and Science and Technology strands seems to be a commonality across the test forms as an area of weakness. However, the results for range-of-knowledge correspondence in particular clearly point to narrow assessment of the full set of GLEs within all strands for each grade assessment reviewed. As a result, the findings on balance-of-knowledge representation should be interpreted with caution.

Our general conclusion regarding the range of content assessed is that DESE should consider substantial modification either to the test items or the GLEs. As specified by the USDE (2004), assessments should align to the full range of the content expectations established by the state. Although Missouri does prioritize some GLEs, it is still the case that the test forms fail to represent a sizeable portion of the GLEs for each grade.

²² These results could be different if items were matched only to the assessed grade's GLEs rather than the total GLEs across the grade span (Grades 3-5 and 6-8).

Suggestions for improving the alignment between the science assessments and Missouri Grade-Level Expectations are discussed in Chapter 6 Summary and Recommendations.

Chapter 6 Summary and Recommendations

HumRRO conducted a review of the MAP to examine content alignment to the Missouri Grade-Level Expectations for Communication Arts-Reading and Writing, Mathematics, and Science. Alignment of assessments and achievement standards to the state academic content standards is a requirement of the No Child Left Behind Act of 2001.

The extent of alignment to the Missouri Grade-Level Expectations differed considerably per content area and grade. The 2010 and 2011 test forms for Mathematics demonstrated the strongest alignment to the GLEs. The Communication Arts test forms displayed the most variability in alignment across grades and alignment criteria. For example, Communication Arts results suggest that the majority of grade-level test forms assess students on a range of the GLEs within content strands/Big Ideas. Furthermore, the test forms assess the major Reading categories with a sufficient number of items for over half of the Big Ideas, although the Writing assessment may warrant review to ensure that these content expectations receive adequate emphasis as well. In comparison, the Communication Arts items tended to cluster around a small number of assessed GLEs, producing unbalanced content coverage. The test blueprints provide guidance on emphasis across strands within a content area and not at the GLE level. As a result, the number of items per strand may meet the test blueprint guidance, but still have unbalanced content coverage across GLEs within a strand. Finally, over 50% of items on the majority of Communication Arts test forms assess students at a lower level of cognitive processing than required in the Missouri Grade-Level Expectations.

Findings for Science require some additional explanation for the following reasons. In matching items to Science GLEs, panel members matched items against the following choices: all strands, concepts, and assessed GLEs for the grade span (Grade Span 3-5 and Grade Span 6-8). As a result, when the panel members then matched the items to the GLEs, they could match an item to any of the grade span's assessed GLEs. For Grade 5, panel members matched the 63 items from the 2010 form and 64 items on the 2011 form to the 149 rated strands, concepts, and assessed GLEs for Grade Span 3-5. If panel members were restricted to matching against only those GLEs assessed for Grade 5, they would have been matching the items to only 56 GLEs. Grade 8 procedures were the same, but for Grade Span 6-8. Approximately 16% of the panelists' ratings matched items to Grade 3 standards, 22% to Grade 4 standards, and 62% to Grade 5 standards for the Grade 5 assessment. Approximately 24% of panelists' ratings matched items to Grade 6 standards, 27% to Grade 7 standards, and 50% to Grade 8 standards on the Grade 8 assessment. All panelists ratings are provided in the Appendices Tables C-13 and C-14.

The decision to have panelists match items to the grade span rather than to a single grade most directly affects the results in range of knowledge. For range of knowledge, only 1 strand was found to be adequately assessed from the possible 32 Science strands across both forms and grade levels. However, matching the 63 and 64

items for Grade 5 forms 2010 and 2011, respectively, to 56 rather than 149 choices provides a far higher likelihood for matching at least 50% of the GLEs within each strand. The same logic holds for matching the 65 and 64 items for Grade 8 forms 2010 and 2011, respectively, to 82 rather than 219 choices.

Categorical concurrence should not be affected since the strands remain constant across all grades.

Depth of knowledge may have been affected, but not to a large extent. The impact would probably be that more items were at the same or higher DOK level as the standards since the items also were being matched to lower grades' GLEs rather than only for Grades 5 and 8 (assuming that there was a tendency for standards to have lower DOK requirements at lower grades). This may have actually increased the number of strands that were determined to be adequately assessed.

Balance of knowledge also may be higher as a result of the items being dispersed over more choices. A quick scan of the data matches for Grades 5 and 8 found that the most frequently selected choices had only one or two items matched and very few choices with more than three items matched.

However, it is not possible to examine the data to match only with the Grade 5 and Grade 8 GLEs. Panelists matched items to standards from the grade span, and we cannot ascertain what standards they would have matched, or if they would have matched standards at all, if only a single grade's standards had been presented. All analyses and results for Science in this report were based on the assumption that each assessment represented a three-grade span.

Table 6.1 provides summary alignment conclusions for each grade level and content area per Webb alignment indicator. This table provides the summary alignment judgments from Tables 3.12, 4.13, and 5.13.

Table 6.1. Summary Degree of Alignment Conclusions per Grade and Content Level for Each Webb Alignment Indicator

Content Area and Grade	2010 Test Form				2011 Test Form			
	Percentage of Strands that Met Webb Criteria				Percentage of Strands that Met Webb Criteria			
	CC	DOK	ROK	BOK	CC	DOK	ROK	BOK
Comm Arts								
3	High	Partial	High	Weak	Partial	Partial	High	Partial
4	Partial	Partial	Partial	Partial	Partial	Weak	Full	Partial
5	Partial	Partial	High	Partial	Partial	Partial	Partial	Partial
6	Partial	Weak	High	Weak	Partial	Partial	Full	Partial
7	Partial	Weak	Full	Partial	Partial	Partial	High	Weak
8	Partial	Partial	High	Partial	Partial	High	Full	Partial
Math								
3	Full	Full	Full	Full	Full	Full	Full	Full
4	Full	Full	Full	Full	Full	Full	Full	High
5	Full	High	Full	Full	Full	Full	Full	Full
6	Full	High	Full	High	Full	High	Full	High
7	Full	Full	Full	Partial	Full	Full	Full	Partial
8	Full	Full	Full	Full	Full	Partial	Full	Partial
Science								
5	High	Full	Weak	Full	Partial	High	Weak	Full
8	Partial	Partial	Weak	High	High	Weak	Weak	High

Note: CC = Categorical Concurrence; DOK = Depth-of-knowledge Consistency; ROK = Range-of-knowledge Correspondence; BOK = Balance-of-knowledge Representation

Based on these results, HumRRO makes the following recommendations to Missouri on ways in which test alignment might be improved. These recommendations focus on the more critical findings. 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 grade-level test and content area are justifiable, meaning the state intentionally chose to emphasize some strands and GLEs over others. In these cases, DESE should consider explicitly including these justifications in test documentation.

Recommendations

Communication Arts

- 1. Consider ways to increase overall content coverage on the assessments, particularly for Writing content expectations (categorical concurrence).** The content expectations composing the Big Ideas on Writing-Process and Writing-Forms/Types that Missouri expects students to know currently appear under-represented (less than 6 items per Big Idea) on the assessments, resulting in a conclusion of ‘partial alignment’ overall. Coverage of major content categories could be increased or explained in several ways: (a) increase number of items [approximately 4-6 selected response (SR) items per Big Idea], (b) if constructed response (CR) items target multiple content areas, provide more explicit description of possible content coverage for these items in test documentation to gain more transparency in demonstrating alignment, (c) consider developing, or modifying, CR items to target additional content, and/or (d) explicitly note in GLEs and test documentation why this Writing content is not assessed at the state level and describe how students are expected to demonstrate this knowledge in other ways.
- 2. Evaluate the cognitive complexity assessed by items relative to the Missouri Grade-Level Expectations for both grade-level test forms (depth-of-knowledge consistency).** With the exception of 2011 form for Grade 8, the panelists reviewing these assessments rated a number of items as less demanding cognitively than the Missouri Grade-Level Expectations. Thus, the assessments may not adequately reflect the rigor of the state standards for some content expectations. This finding is not uncommon among large-scale assessments. However, such a circumstance also is not an inevitable consequence of standardized testing. The number of adjacent ratings (DOK of 1 vs. 2) given by reviewers suggests only moderate discrepancy between items and GLEs. Thus, increasing cognitive complexity may require minor modifications to items.
- 3. Review the ratio of items assigned to assessed GLEs within each Big Idea for all grades to evaluate content emphasis on the assessment (balance-of-knowledge representation).** While the majority of grade-level Communication Arts test forms assessed a range of GLEs per strand, the distribution of items among these GLEs appears unbalanced. In other words, reviewer ratings suggest that a number of items cluster around one to two GLEs. This type of problem can be remedied in several ways: (a) increase the number of items assigned to GLEs with low emphasis, (b) redistribute existing points (requiring some new item construction or modification) among GLEs more evenly, or (c) provide more explicit justification for uneven content emphasis. The solution chosen depends on various constraints (usually time and money) that exist for DESE and for the test vendor.

Mathematics

1. **Review item assignment to content expectations for increased alignment to assessed GLEs within each strand for Grades 7 and 8 (balance-of-knowledge representation).** The majority of results for the grade-level Math test forms indicate high alignment to the GLEs. One area that DESE may wish to review is the item distribution among GLEs, particularly for assessment of the Number and Operations strand and the Data and Probability strand on the Grade 7 and 8 test forms. The Math test forms demonstrated alignment to a broad range of GLEs; thus, item clustering around several GLEs within these strands is the most likely explanation for reduced alignment in these cases. As with Communication Arts, three options may be considered for increasing balanced alignment (see above). The second option of redistributing points/items among GLEs may be the most practical option, even between strands, since about half of total items were matched by reviewers to GLEs within the Numbers and Operations and Algebraic Relationships strands.

Science

1. **Review the breadth of content covered on the 2011 test forms for Grades 5 and 8 to increase alignment to the Missouri Grade-Level expectations (categorical concurrence and range-of-knowledge correspondence).** The results on the test forms for Grades 5 and 8 indicate that these assessments do not meet the minimum criteria for several alignment measures when compared to the GLEs to be assessed for the grade spans (Grades 3-5 and Grades 6-8). The most critical issue pertains to the small percentage of Science GLEs within the grade span assessed by each grade-level test form, which is evident from the range-of-knowledge representation results. Thus, the assessments do not adequately “cover the full range of content specified in the State’s academic content standards” (USDE, 2004, p.41) for the grade span. If the state considers all of these content expectations for the grade span important for students to know in order to demonstrate mastery of grade span Science concepts, then the MAP should assess a larger proportion of the grade span content expectations.

This issue may be a result of combining for this study the GLEs for all grades within the grade spans (3-5 and 6-8) rather than examining the alignment of the Grades 5 and 8 Science assessment only to the that grade’s science GLEs. Examining only the grade-specific GLEs with the grade-level assessment provides a far higher likelihood of matching at least 50% of the GLEs within each strand to a test item to meet the Webb criterion for range of knowledge.

Related to content coverage at a broader level, reviewers found that some

strands were assessed by fewer than six items²³. Specifically, the 2011 Grade 5 test form and the 2010 Grade 8 test form did not meet the minimum criterion for adequate assessment of the strands Force and Motion, Universe, and Science and Technology. This outcome also is a symptom of an unbalanced ratio of test items to standards; however, the alignment issue is less critical for two reasons: (a) the mean number of items matched is very close to six in each case, and (b) the content emphasis on these assessments is comparable to the test blueprint. While some researchers argue that a minimum of six items is arbitrary, an assessment should include a sufficient number of items for accurate assessment of what students know to produce valid scores.

- 2. Evaluate the cognitive complexity assessed by items relative to the Missouri Grade-Level Expectations on the Grade 8 test forms (depth-of-knowledge consistency).** Reviewer ratings of item DOK for both Grades 5 and 8 suggest that the test forms assess a lower level of cognitive complexity overall than required by the content expectations²⁴. The results for the Grade 8 test forms, in particular, indicate a more marked discrepancy (majority of items assessed as DOK level 1, while many GLEs rated as DOK level 2²⁵). Two issues deserve consideration. First, while the magnitude of discrepancy between items and GLEs is low (DOK of 1 vs. 2), the number of items falling below the cognitive complexity level expected in corresponding GLEs is high. As a result, students rarely must demonstrate knowledge at the same level as the content standards. Second, and somewhat surprisingly, the performance expected of students in the majority of GLEs qualifies as lower-order cognitive processing. Relatively few GLEs expect students to master Science concepts at a higher level requiring complex reasoning (DOK level 3)²⁶. Science concepts often involve greater *difficulty* due to the cumulative nature of Science knowledge acquisition. While difficulty and complex cognitive processing are correlated, *difficult* concepts requiring more prerequisite knowledge do not necessarily involve in-depth cognitive processing. DESE and the test developer may wish to review the GLEs in addition to the test forms to further examine whether the test items expect students to demonstrate comprehension and application of science concepts at the cognitive complexity level required of the students by the GLEs.

²³ Since strands remained the same for across the grade spans, this criterion should not have been impacted by matching to all GLEs across the grade span.

²⁴ The results for this criterion should not have been adversely impacted by examining items to all GLEs within each grade span.

²⁵ The HumRRO Alignment Panel rated 48 of 82 Grade 8 GLEs at DOK level 2 while the DESE Standards Writing Committee rated 49 GLEs at DOK level 2. There were three differences between the ratings by the Panel and the Committee. One GLE was rated one level higher by the Panel and 2 GLEs were rated one level higher by the Committee.

²⁶ Both the Panel and the Committee rated 4 GLEs at DOK level 3 and 1 GLE at DOK level 4.

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*,
- 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 mathematics and science education*. Washington, D.C.: Council of Chief State Schools Officers.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states. (Research Monograph 18)*. Madison, WI: National Institute for Science Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852)

