

Missouri Assessment Program-Alternate (MAP-A) Alignment Review for Science: Technical Report

**Leslie R. Taylor
Hilary L. Campbell
Rebecca L. Norman Dvorak
Richard C. Deatz
Lisa E. Koger
Milton E. Koger
D. E. (Sunny) Becker
Arthur A. Thacker**

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

Prepared under: Contract No: C308004001-002

December 18, 2009

Missouri Assessment Program-Alternate (MAP-A) Alignment Review for Science: Technical Report

**Leslie R. Taylor
Hilary L. Campbell
Rebecca L. Norman Dvorak
Richard C. Deatz
Lisa E. Koger
Milton E. Koger
D. E. (Sunny) Becker
Arthur A. Thacker**

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

Prepared under: Contract No: C308004001-002

December 18, 2009

EXECUTIVE SUMMARY

Purpose and Scope of Work

The Missouri Department of Elementary and Secondary Education (DESE) requested an external independent alignment study (evaluation/analysis) of the Missouri Assessment Program-Alternate (MAP-A) in Science for students with significant cognitive disabilities. Specifically, DESE wanted an evaluation of the alignment between the MAP-A portfolio assessment, the extended content standards (or Alternate Grade-Level Expectations¹), and the Missouri Show-Me Standards². Missouri uses the MAP-A portfolio assessment in the federal and state accountability programs. DESE awarded Human Resources Research Organization (HumRRO) the contract to conduct this alignment study, and work began on February 1, 2009.

DESE requested the alignment study to meet both state and federal requirements. The federal requirement of the U.S. Department of Education (USDE) stems from the *No Child Left Behind Act* (NCLB) of 2001. NCLB challenges each state to establish a coherent assessment system based on solid academic standards. This law calls for states to provide independent evidence of the validity of its 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 these assessments measure student knowledge in the same manner and at the same level of complexity as expected 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.

Alternate assessments are included in the federal requirements. The federal government has established regulations for students with significant cognitive disabilities in the calculation of school and district AYP determinations, often referred to as the “1% rule” (U.S. Department of Education, 2005). This rule allows the state to accommodate students with significant cognitive disabilities in

¹ Missouri Alternate Grade-Level Expectations can be found at: http://www.dese.mo.gov/divimprove/assess/mapa_resources.html

² Missouri Show-Me Standards can be found at: <http://www.dese.mo.gov/standards/>

its AYP calculations by setting different performance expectations for up to 1% of the student population. As a result, states can develop alternate content standards (often referred to as extended standards), achievement standards, and assessments designed to more fairly demonstrate the knowledge of these students. However, the content on which these students are assessed must be academic, and the achievement of these students must continue to reflect challenging academic goals. As such, states must show that the extended standards and alternate achievement standards for these students link to the grade-level expectations, although the breadth and depth of these expectations can be reduced (USDE, 2005).

Methodology

HumRRO convened panels of Missouri educators and national content experts to review the MAP-A portfolio assessments. These panelists included current and former teachers, administrators, and curriculum specialists/district coordinators.

Seven panelists met to review the Science portfolios relative to the Missouri Alternate Grade-Level Expectations (AGLEs). The panel consisted of six in-state Missouri panelists and one out-of-state panelist. Each panelist evaluated portfolios for Grades 5, 8, and 11 in Science.

HumRRO used the Links for Academic Learning alignment method (referred to as the LAL method in this report) developed by the National Alternate Assessment Center (NAAC) to conduct the reviews and analyze the results (Flowers, Wakeman, Browder, & Karvonen, 2007). This method requires panelists to rate the content standards and assessments on multiple dimensions. Ratings are then analyzed and interpreted based on seven criteria. These criteria are listed below (adapted from Flowers et al., 2007):

LAL Criterion 1: Academic - The content is academic and includes the major domains/strands of the content area as reflected in state and national standards (e.g., mathematics, reading, science).

LAL Criterion 2: Age Appropriate - The content is referenced to the student's assigned grade level (based on chronological age).

LAL Criterion 3: Standards Fidelity

- a. Content Centrality** - The target content maintains fidelity with the content of the original grade-level standards.
- b. Performance Centrality** - The focus of achievement maintains fidelity with the specified performance in the grade-level standards.

LAL Criterion 4: Content Coverage (Webb alignment indicators) - The content differs from grade level in range, balance, and depth of knowledge

(DOK), but matches high expectations set for students with significant cognitive disabilities.

LAL Criterion 5: Content Differentiation - There is some differentiation in content across grade levels or grade bands.

LAL Criterion 6: Achievement - The expected achievement for students is for the students to show learning of grade referenced academic content.

LAL Criterion 7: Performance Accuracy - The potential barriers to demonstrating what students know and can do are minimized in the assessment to increase measurement accuracy of student performance.

Under LAL Criterion 4 above, we refer to the Webb alignment indicators.” Dr. Norman Webb (2005) developed an alignment procedure involving an evaluation of the assessment to the content standards using four statistics. These statistics indicate how well an assessment covers the content standards in terms of content breadth and depth. Webb’s method generally has been applied to regular general education assessments, and some special education researchers (i.e., Flowers et al., 2007) consider this approach to be limited as a primary alignment method for alternate assessments. However, the Webb alignment indicators provide important information regarding content coverage. Thus, the LAL method includes the following Webb alignment indicators:

- (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., grade-level expectation) assessed within each strand.
- (3) Balance-of-knowledge representation – provides a statistical index reflecting the distribution of assessed content within each strand, meaning how evenly the content is assessed.
- (4) Depth-of-knowledge (DOK) consistency – compares the cognitive complexity ratings of the items with the complexity ratings of each content standard.

The outcomes of the analyses on the LAL criteria and Webb alignment indicators are evaluated against decision rules to judge their acceptability. However, because the MAP-A is a portfolio assessment, with only four entries per student (two for science, each designed to assess two content strands) to assess a large number of Alternate Grade-Level Expectations (AGLEs), we can expect Webb’s Indicators 1-3 from above not to be met. The criterion for meeting the categorical concurrence requirement would be that the assessment include 6 items per content strand (e.g. Matter and Energy), so even if all four entries assessed a single strand (which is not possible since each entry is designed to

assess one content and one process strand), the criterion could still not be met. Similarly, the requirement for acceptable range-of-knowledge representation is that at least one item on the assessment relate to 50% of the indicated standards. Because there are so many science AGLEs per strand, this criterion could also not be achieved. Finally, the requirement for balance-of-knowledge correspondence is also inappropriate. Each science portfolio is expected to include two content strands (there are two assessed at each grade) and two process strands. The typical portfolio will contain one entry for each assessed strand and will also assess each of the two process strands. The design of the system typically results in a portfolio assessing one of each of the indicated strands—and consequently only one of the AGLEs associated with each strand. The balance index is inappropriate because it is misleading. Since each portfolio is likely to have only 1 AGLE per strand, the index would be high (same number of AGLEs per strand), but this would not inform us about the distribution of the AGLEs by strand.

Webb’s DOK consistency requirement is appropriate for this study. The match between the DOK of the AGLEs and the DOK indicated by the portfolio entries can be ascertained and reported. For the other indicators, HumRRO chose to describe the distribution of the AGLEs across portfolios to give DESE information about which AGLEs were assessed within strands and whether certain AGLEs were favored while others were avoided. We reasoned that if there was a reasonable distribution of AGLEs across portfolios within a grade, a student might have the opportunity to receive instruction across several AGLEs within each strand during the assessment grade and previously. It is still possible that a single student might be instructed on the same 4 AGLEs for multiple years, but if there is a wide distribution of AGLEs assessed among the sample of portfolios and if no AGLE dominates the assessments, this seems unlikely. The ratings included in this report for Webb’s Criteria 1-3 (noted above) do not represent acceptability using Webb’s interpretations.

Summary Alignment Results

Key Findings and Conclusions

The results of the alignment reviews provide positive support for the content validity of the MAP-A assessment based on several outcomes. First, the majority of panelists found all of the grade-level AGLEs for science to be linked adequately to the full Missouri Show-Me Standards in content breadth and depth. Second, nearly all MAP-A portfolio entries across grades were rated as matched to AGLEs. Third, panelists determined that the AGLEs and assessments are accessible to a wide range students with various physical and cognitive disabilities. Finally, the alignment review of the achievement standards to portfolio entries suggests that the Science MAP-A assessments are designed to reflect DOK levels similar to the DOK level indicated by the AGLEs.

As with most alignment reviews, the findings also point to some areas where content and performance alignment could be strengthened over time.

AGLEs to Missouri Show-Me Standards

Table 1 displays the summary conclusions regarding content alignment between the AGLEs and Missouri Show-Me Standards for Science. These judgments are based on whether the AGLEs achieved acceptable levels of linkage with the full content standards for each grade level. The minimum level for each of the criteria in Table 1 is 90%.

- High linkage - most of standards are acceptable (at least 90%)
- Partial linkage - some standards are acceptable (50-89%)
- Weak linkage - few to no standards are acceptable (less than 50%)

Table 1. Summary Conclusions on Alignment of AGLEs to Missouri Show-Me Standards for Science on LAL Criteria 2, 3, and 5

	Criterion 2	Criterion 3		Criterion 5
Grade	Age Appropriate	Content Centrality	Performance Centrality	Content Differentiation
	Is content referenced to student's assigned grade level?	Do the extended standards link to the target content in the grade-level standards?	Does the performance of the extended standards link to expectations of the grade level standards?	Do the extended standards show appropriate increases between grade levels?
5	High	High	High	High
8	High	High	High	High
11	Partial	High	High	High

The content alignment conclusions in Table 1 indicate that the grade-level AGLEs link well to the Missouri Show-Me Standards across grades. In addition, this content is well differentiated, or vertically aligned, between grades overall.

Table 2 displays the overall conclusions pertaining to LAL Criterion 7 - Performance Accuracy (content accessibility) for the AGLEs. For this criterion, conclusions reflect overall judgments of acceptability based on the following categories³.

³ Adapted from universal design ratings used by the National Center on Educational Outcomes (NCEO). See Thompson et al. (2005).

- Excellent - all standards are acceptable
- Good - most standards are acceptable (at least 90%)
- Acceptable - many standards are acceptable (70%-90%)
- Questionable - few standards are acceptable (less than 70%)

Table 2. Summary Conclusions on Performance Accuracy (LAL Criterion 7) of Science AGLEs

Criterion 7		
Grade	Performance Accuracy (Potential Barriers to Accessibility)	
	Is the content appropriate for students at different levels of communication?	Is the content accessible to different disability groups?
5	Questionable	Good
8	Questionable	Good
11	Questionable	Good

The conclusions on Performance Accuracy clearly are disparate. Although seemingly in conflict, the two ratings on access address quite different aspects. The conclusions on communication reflect panelists' judgments that students with limited symbolic ability may have difficulty comprehending or demonstrating the content in some AGLEs. In contrast, panelists felt that students with varying physical or cognitive impairments (i.e., difficulty with instructions, attention, sensory integration) can typically access and demonstrate this knowledge.

MAP-A Tasks to AGLEs

Table 3 provides summary conclusions on the alignment of the MAP-A assessments to AGLEs. The conclusions are based on the following criteria:

- High linkage - most of tasks are acceptable (at least 90%)
- Partial linkage - some tasks are acceptable (50-89%)
- Weak linkage - few to no tasks are acceptable (less than 50%)

Table 3. Summary Conclusions on Alignment of Science MAP-A Portfolios to AGLEs for LAL Criteria 1, 2, 3, 4, and 5

	Criterion 1	Criterion2	Criterion 3		Criterion 4		Criterion 5
Grade	Academic Content	Age Appropriate	Content Centrality	Performance Centrality	Content Coverage		Content Differentiation
	Are students assessed on academic content?	Is task content referenced to student’s assigned grade level?	Do tasks link to the target content in the AGLEs?	Does the performance of task link to expectations of the AGLEs?	Do the tasks assess students at the appropriate breadth of knowledge? ^a	Do the tasks assess students at the appropriate depth-of-knowledge? ^b	Do the assessments show appropriate increases between grade levels?
5	High	High	Partial	High	Weak	Partial	Weak
8	High	Partial	Partial	High	Weak	Partial	Weak
11	High	Partial	Partial	High	Weak	High	Partial

^a These conclusions are based on a summary judgment across the Webb statistics of Categorical Concurrence, Range of Knowledge, and Balance of Knowledge, which may be inappropriate for this assessment.

^b These conclusions are based on the results from the DOK consistency analyses.

As Table 3 illustrates, the 2009 MAP-A assessments linked well to the AGLEs on the majority of dimensions. Under LAL Criterion 4, indications of alignment were restricted by the numbers of portfolio entries compared to AGLEs as described above. These criteria may not be appropriate for this assessment, but it does seem clear that there are more AGLEs than can be adequately assessed from a single portfolio. On DOK assessed, panelists determined that at least half of the Grades 5 and 8 tasks under Scientific Inquiry assessed students at a lower level of cognitive complexity than expected in the AGLEs. All other strands for all grades were represented by at least 50% of portfolio entries at or above the DOK level indicated by the AGLEs. Regarding LAL Criterion 5, Content Differentiation, panelists did find the content between each grade-level assessment to demonstrate at least some content differentiation, although panelists felt that the content in the Grade 11 assessments showed the clearest increases in breadth, depth, and new knowledge. Panelists typically rated aspects of content differentiation for Grades 3-8 as showing limited or no content differentiation by grade level.

Table 4 includes results related to Criteria 6 and 7 of the LAL method. These rating questions asked panelists to determine whether the assessment tasks are designed in such a way that students can demonstrate knowledge at various levels of functioning and ability. Ratings in this case are based on evaluations of accessibility, rather than on content alignment⁴.

- Excellent - all tasks are acceptable
- Good - most tasks are acceptable (at least 90%)
- Acceptable - many tasks are acceptable (70%-90%)
- Questionable - few tasks are acceptable (less than 70%)

Table 4. Summary Conclusions on Accessibility (LAL Criteria 6 and 7) of MAP-A Portfolios

	Criterion 6	Criterion 7		
Grade	Achievement	Performance Accuracy (Potential Barriers)		
	Does the assessment allow for accurate inference about student learning?	What level of symbolic communication does task require?	Is task accessible to different disability groups?	Can task be modified/supports provided without changing meaning or difficulty?
5	Questionable	Questionable	Excellent	Excellent
8	Questionable	Questionable	Good	Excellent
11	Questionable	Questionable	Good	Excellent

⁴ Alignment refers to overlap in content expectations. In this case, the goal is not to measure the test against the content expectations but to evaluate the level of accessibility.

The most noticeable issue regarding accessibility for the MAP-A science portfolio assessments concerns the measurement of achievement (LAL Criterion 6). Panelists indicated difficulty in being able to make inferences about student achievement regarding degree of new learning, generalizability, and program indicators. Panelists were concerned that new learning might be difficult to measure since the content of the portfolio could change radically from year to year and because assessed strands changed based on grade level. Panelists were also concerned that the portfolio entries might not generalize across multiple AGLEs within a strand. Finally, the panelists were concerned that the portfolios did not provide a strong indication of program quality.

Results on LAL Criterion 7 were mostly positive, with the exception of level of symbolic communication. It should be noted that the panelists were rating portfolios designed for particular students—rather than the system for creating portfolios itself. It might not be surprising that most panelists judged that creating these particular portfolio entries would require symbolic communication beyond pre-symbolic. All grades showed otherwise acceptable results for this criterion.

Recommendations

HumRRO makes the following recommendations to strengthen the linkage between the components of the Missouri alternate assessment system.

AGLEs for Science

- (1) **Review the access points for the AGLEs at each grade level.** For each grade level, panelists identified some AGLEs that may limit access only to those students with higher symbolic abilities, thus excluding a portion of students from the assessment system. Reviewing the AGLEs may involve additional bias reviews to modify the current expectations; or, additional explanation (e.g., content limitations, examples) within the MAP-A Test Specifications document may be sufficient to better illustrate how teachers might make these content expectations more appropriate for students with lower symbolic abilities.

MAP-A Portfolio Tasks

- (1) **Review performance tasks for age appropriateness for Grades 8 and 11.** A portion of the portfolio tasks (12-16%) for these grades was judged by panelists to be age inappropriate. While not a large percentage, this may indicate the need for additional training of teachers or that clearer instructions be provided for preparing the portfolio tasks.
- (2) **Review performance tasks for content centrality for all grades.** A portion of the portfolio tasks (14-22%) was judged by panelists to

be weakly linked or not linked to the AGLEs. This may indicate the need for additional training of teachers or clearer instructions for preparing the portfolio tasks to target specific AGLEs.

- (3) ***Consider the implications of using a portfolio system to assess standards with so many AGLEs per strand.*** Webb's alignment criteria related to breadth of content knowledge were inappropriate for this assessment because of the small numbers of portfolio entries compared to the relatively large numbers of AGLEs. There are simply too many AGLEs to adequately measure with the limited number of portfolio entries.

- (4) ***Consider the inferences expected to be made from science MAP-A scores by schools and teachers in relation to the test design.*** Panelists indicated concerns that the science MAP-A scores might not provide adequate information about students' new learning, that they might not generalize beyond the specific content of the portfolio task, and that the scores might not reflect program quality. Panelists were concerned about the impact of teachers on test scores due to their capacity to create appropriate portfolio tasks.

MISSOURI ASSESSMENT PROGRAM-ALTERNATE (MAP-A) ALIGNMENT REVIEW: TECHNICAL REPORT

TABLE OF CONTENTS

List of Tables	xii
Chapter 1 Introduction	1
Alignment Requirements for State Assessment Systems.....	1
Structure of Missouri's Alternate Assessment System	2
Chapter 2 Alignment Study Design and Methodology	4
Alignment of Assessments and Standards on Content and Accessibility	4
<i>Links for Academic Learning (LAL) Alignment Method</i>	5
<i>Panelists</i>	7
<i>Materials</i>	8
<i>Procedures</i>	8
Chapter 3 Results for MAP-A Science	11
Alignment of Science AGLEs to Missouri Show-Me Standards.....	11
Alignment of Science Portfolio Assessments to AGLEs	20
Inter-Rater Agreement Results	39
References	42
Appendix A Alignment Results per Assessment	A-1
Appendix B Workshop Instructions	B-1

List of Tables

Table 1. Summary Conclusions on Alignment of AGLEs to Missouri Show-Me Standards for Science on LAL Criteria 2, 3, and 5.....	vi
Table 2. Summary Conclusions on Performance Accuracy (LAL Criterion 7) of Science AGLEs.....	vii
Table 3. Summary Conclusions on Alignment of Science MAP-A Portfolios to AGLEs for LAL Criteria 1, 2, 3, 4, and 5	viii
Table 4. Summary Conclusions on Accessibility (LAL Criteria 6 and 7) of MAP-A Portfolios.....	ix
Table 2.1 Professional and Demographic Characteristics of Science MAP-A Alignment Panelists	8
Table 3.1. Mean Number of Science AGLEs Rated as Academic by Panelists ..	11
Table 3.2. Science AGLEs at Various Levels of Age Appropriateness	12
Table 3.3. Mean Number of AGLEs at Various Levels of Content Centrality	14

Table 3.4. Mean Number of AGLEs at Various Levels of Performance Centrality	15
Table 3.5. Consensus Ratings on Content Differentiation between Grades MAP-A AGLEs for Science	16
Table 3.6. Mean Number of AGLEs Rated at Each Level of Symbolic Communication	18
Table 3.7. Science AGLEs Rated as Accessible to All Students	19
Table 3.8. Science AGLEs Rated as Amenable to Accommodations for All Students	19
Table 3.9. Mean Number of Science MAP-A Portfolio Tasks Rated as Academic by Panelists	21
Table 3.10. Science MAP-A Performance Tasks at Various Levels of Age Appropriateness	22
Table 3.11. Mean Percent of Tasks Linked to AGLEs	22
Table 3.12. Mean Percent of Tasks at Various Levels of Content Centrality	23
Table 3.13. Mean Percent of Tasks at Various Levels of Performance Centrality	24
Table 3.14. Summary of Breadth Across Content Categories Results for MAP-A Science Portfolios by Grade Level	27
Table 3.15. DOK Consistency for Science MAP-A, Grade 5: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives	29
Table 3.16. DOK Consistency for Science MAP-A, Grade 8: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives	30
Table 3.17. DOK Consistency for Science MAP-A, Grade 11: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives	31
Table 3.18. Mean Percentage of Tasks at Each DOK Level	31
Table 3.19. Consensus Ratings on Content Differentiation between Grades MAP-A Assessments for Science	34
Table 3.20. Degree of Inference Evident on Student Learning in Science MAP-A Assessments	36
Table 3.21. Mean Number of Tasks Rated at Each Level of Symbolic Communication	38
Table 3.22. MAP-A Science Performance Tasks Rated as Amenable to Accommodations for All Students	39
Table 3.23. Pairwise Comparisons on AGLE Ratings per Grade Level	40
Table 3.24. Pairwise Comparisons on Portfolio Ratings per Grade Level	41
Table A-1. Content Representation for Science MAP-A: Number of Portfolio Entries per Strand	A-3
Table A-2. Depth-of-Knowledge Consistency for Science MAP-A, Grade 5: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives	A-4
Table A-3. Depth-of-Knowledge Consistency for Science MAP-A, Grade 8: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives	A-5

Table A-4. Depth-of-Knowledge Consistency for Science MAP-A, Grade 11:
Mean Percent of Performance Tasks with DOK Below, At, and Above
DOK Level of Objectives..... A-6

Table A-5. Summary of Portfolio Entry-to-AGLE Distribution Results for MAP-A
Science for Grade 5..... A-7

Table A-6. Summary of Portfolio Entry-to-AGLE Distribution Results for MAP-A
Science for Grade 8..... A-7

Table A-7. Summary of Portfolio Entry-to-AGLE Distribution Results for MAP-A
Science for Grade 11..... A-8

MISSOURI ASSESSMENT PROGRAM-ALTERNATE (MAP-A) ALIGNMENT REVIEW FOR SCIENCE: 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-Alternate (MAP-A) in Science for students with significant cognitive disabilities. Specifically, DESE wanted an evaluation of the alignment between the MAP-A portfolio assessment, the extended content standards (or Alternate Grade-Level Expectations⁵), and the Missouri Show-Me Standards⁶. Missouri uses the MAP-A portfolio assessment in the federal and state accountability programs. DESE awarded Human Resources Research Organization (HumRRO) the contract to conduct this alignment study, and work began on February 1, 2009.

Alignment Requirements for State Assessment Systems

DESE requested the alignment study to meet both state and federal requirements. The federal requirement of the U.S. Department of Education (USDE) stems from the *No Child Left Behind Act* (NCLB) of 2001. NCLB challenges each state to establish a coherent assessment system based on solid academic standards. This law calls for states to provide independent evidence of the validity of its 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 a state's assessment system. Alignment results should demonstrate that the assessments represent the full range of content standards, and that these assessments measure student knowledge in the same manner and at the same level of complexity as expected 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.

Alternate assessments are included in the federal requirements. The federal government has established regulations for students with significant cognitive disabilities in the calculation of school and district AYP determinations, often referred to as the "1% rule" (U.S. Department of Education, 2005). This rule

⁵ Missouri Alternate Grade-Level Expectations can be found at: http://www.dese.mo.gov/divimprove/assess/mapa_resources.html

⁶ Missouri Show-Me Standards can be found at: <http://www.dese.mo.gov/standards/>

allows the state to accommodate students with significant cognitive disabilities in its AYP calculations by setting different performance expectations for up to 1% of the student population. As a result, states can develop alternate content standards (often referred to as extended standards), achievement standards, and assessments designed to more fairly demonstrate the knowledge of these students. However, the content on which these students are assessed must be academic, and the achievement of these students must continue to reflect challenging academic goals. As such, states must show that the extended standards and alternate achievement standards for these students link to the state standards, although the breadth and depth of these expectations can be reduced (USDE, 2005).

Structure of Missouri's Alternate Assessment System

Missouri chose to develop a portfolio system for those students with the most significant cognitive disabilities. A portfolio assessment is unique to each student and based on the student's Individualized Education Plan (IEP), as opposed to a fully standardized assessment with common items or tasks across students. As part of this alternate assessment system, Missouri also constructed Alternate Grade-Level Expectations (AGLEs) on which the portfolio assessments must be based. Per the federal requirement, these content expectations correspond with the Missouri Show-Me Standards, although reduced in breadth and depth.

For Science, portfolios are assessed at Grades 5, 8, and 11, but each might include content learned in prior grades. The Science MAP-A assesses six content strands and two process strands. The design of the assessment is summarized in Figure 1.1 below. For each student, teachers submit two entries with each grade-level assessment. For Science, each entry is expected to address two strands, one content strand and one process strand.

Design		
Content Area	Title of Strand	Grade Focus
Science SCI Process Strands	Scientific Inquiry (IN)	5, 8, & 11
	Impact of Science, Technology, and Human Activity (ST)	5, 8, & 11
Science SCI Content Strands	Characteristics and Interactions of Living Organisms (LO)	5
	Changes in Ecosystems and Interactions of Organisms with Their Environments (EC)	5
	Properties and Principles of Matter and Energy (ME)	8
	Properties and Principles of Force and Motion (FM)	8
	Processes and Interactions of the Earth's Systems (ES)	11
	Composition and Structure of the Universe and the Motion of the Objects within It (UN)	11

Figure 1.1 Design of the Science MAP-A Assessment⁷

Organization and Contents of the Report

This report contains three chapters. Chapter 2 describes the alignment method and test review details, including panelist characteristics, materials, and procedures. Chapter 3 provides alignment results for Science on the alignment of the AGLEs to the full Missouri Show-Me Standards, alignment of the MAP-A portfolio tasks to the AGLEs, and accessibility of AGLEs and MAP-A portfolios to those students who take this assessment.

Additional information is provided in the appendices of this report. Appendix A contains tables providing more detail on the content alignment results per content area and grade-level test form. Appendix B includes task descriptions and definitions of terms used by panelists who provided the ratings on which this report is based.

⁷ Figure taken from DESE's MAP-A informational Power Point presentation, prepared in August 2009.

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 study.

Alignment of Assessments and Standards on Content and Accessibility

The term *alignment* in this context refers to the degree of consistency evident in instruction and measurement of the state's academic content standards. School curricula should include appropriate content 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. In essence, all alignment evaluations link to the state 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 alignment methods are currently in use for general education and alternate assessments. Most of these methods involve rating various aspects of test items or performance tasks relative to the content standards. Generally, education experts serve as panelists who review and rate the assessments on several measures of content breadth and depth to determine the extent of alignment.

Alignment studies of alternate assessments often require review of additional aspects of alignment unique to those assessments. These dimensions include: (a) accessibility of the assessment system to students with a variety of disabilities, (b) the extent to which test content is academic, and (c) the extent to which alternate content standards are linked with the state's general academic

standards. Alternate assessments differ from general state assessments in form and structure; thus, an alignment methodology must be responsive to these differences.

Links for Academic Learning (LAL) Alignment Method

For the current alignment study, HumRRO applied the Links for Academic Learning alignment method (LAL) developed by the National Alternate Assessment Center to conduct the content alignment reviews and analyze the results (Flowers, Wakeman, Browder, & Karvonen, 2007). This method requires panelists to rate the content standards and assessments on multiple dimensions. Ratings are then analyzed and interpreted based on the following seven criteria (adapted from Flowers et al, 2007):

LAL Criterion 1: Academic - The content is academic and includes the major domains/strands of the content area as reflected in state and national standards (e.g., Mathematics, Reading, Science).

LAL Criterion 2: Age Appropriate - The content is referenced to the student's assigned grade level (based on chronological age).

LAL Criterion 3: Standards Fidelity

a. Content Centrality - The target content maintains fidelity with the content of the original grade-level standards.

b. Performance Centrality - The focus of achievement maintains fidelity with the specified performance in the grade-level standards.

LAL Criterion 4: Content Coverage (Webb alignment indicators) - The content differs from grade level in range, balance, and depth of knowledge (DOK), but matches high expectations set for students with significant cognitive disabilities.

LAL Criterion 5: Content Differentiation - There is some differentiation in content across grade levels or grade bands.

LAL Criterion 6: Achievement - The expected achievement for students is for the students to show learning of grade referenced academic content.

LAL Criterion 7: Performance Accuracy - The potential barriers to demonstrating what students know and can do are minimized in the assessment to increase measurement accuracy of student performance.

The LAL method is appropriate for alignment of multiple assessment formats (i.e., test forms, performance tasks, performance portfolios, and checklists). The method allows for comparison of the assessment to extended standards, as well as alignment of extended standards to full state content standards. In addition, the LAL method includes steps to evaluate the

accessibility of the extended standards and the assessment to students. The review of assessments to standards, such as the MAP-A portfolio assessment to the AGLEs, includes all of the LAL Criteria 1 through 7. The LAL Criteria 1, 2, 3, 5, and 7 apply to a review of extended standards, while all seven criteria apply to a review of the assessment.

Under LAL Criterion 4 above, we refer to the Webb alignment indicators.” Dr. Norman Webb (2005) developed an alignment procedure involving an evaluation of the assessment to the content standards using four statistics. These statistics indicate how well an assessment covers the content standards in terms of content breadth and depth. Webb’s method generally has been applied to regular general education assessments, and some special education researchers (i.e., Flowers et al., 2007) consider this approach to be limited as a primary alignment method for alternate assessments. However, the Webb alignment indicators are still informative regarding content coverage. Thus, the LAL method includes the Webb alignment indicators. These 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., grade-level expectation) assessed within each strand.
- (3) Balance-of-knowledge representation – provides a statistical index reflecting the distribution of assessed content within each strand, representing 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.

The outcomes of the analyses on the LAL criteria and Webb alignment indicators are evaluated against decision rules to judge their acceptability. However, because the MAP-A is a portfolio assessment, with only four entries per student (two for Science, each designed to assess two content strands), we can expect that Webb’s Indicators 1-3 referenced above will not be met. The criterion for meeting the categorical concurrence requirement would be that the assessment includes six items per content strand (e.g. Matter and Energy), so even if all four entries assessed a single strand (which is not possible since each entry is designed to assess one content and one process strand), the criterion could still not be met. Similarly, the requirement for acceptable range-of-knowledge representation is that at least one item on the assessment relate to 50% of the indicated standards. Because there are so many Science AGLEs per strand, this criterion could also not be achieved. Finally, the requirement for balance-of-knowledge correspondence is also inappropriate. Each Science portfolio is expected to include two content strands (there are two assessed at

each grade) and two process strands. The typical portfolio will contain one entry for each assessed strand and will also assess each of the two process strands. The design of the system typically results in a portfolio assessing one of each of the indicated strands—and consequently only one of the AGLEs associated with each strand. The balance index is inappropriate because it is misleading. Since each portfolio is likely to have only 1 AGLE per strand, the index would be high (same number of AGLEs per strand), but this would not inform us about the distribution of the AGLEs by strand.

Webb’s DOK consistency requirement is appropriate for this study. The match between the DOK of the AGLEs and the DOK indicated by the portfolio entries can be ascertained and reported. For the other indicators, HumRRO chose to describe the distribution of the AGLEs across portfolios to give DESE information about which AGLEs were assessed within strands and whether certain AGLEs were favored while others were avoided. We determined that if there was a reasonable distribution of AGLEs across portfolios within a grade (e.g. Grade 5) that a student might have the opportunity to receive instruction across several AGLEs within each strand during the assessment grade and in prior grades (e.g. from Grades 3-4). It is still possible that a single student might be instructed on the same four AGLEs for multiple years, but if there is a wide distribution of AGLEs assessed among the sample of portfolios and if no AGLE dominates the assessments, this seems unlikely. The ratings included in this report for Webb’s Criteria 1-3 above do not represent acceptability using Webb’s interpretations.

For the MAP-A alignment review using the LAL method, Missouri and out-of-state educators performed multiple ratings to carry out the two primary alignment tasks: (a) comparison of the grade-level AGLEs to the Science Missouri Show-Me Standards, and (b) comparison of the Science MAP-A performance tasks (per grade test) to the AGLEs. These tasks served as the basis for the content alignment evaluation of the AGLEs and assessments relative to the full Missouri Show-Me Standards, as well as the content accessibility evaluation of the AGLEs and assessments relative to the population of students for whom the alternate assessment was designed.

Panelists

HumRRO convened a panel of Missouri educators and national content experts to review the Science MAP-A portfolio assessments. These panelists included current and former teachers, administrators, and curriculum specialists/district coordinators. The panel consisted of seven members; six in-state Missouri panelists and one out-of-state panelist. Each reviewer evaluated portfolios for all grade levels under review. Table 2.1 presents the characteristics of the panelists.

Table 2.1 Professional and Demographic Characteristics of Science MAP-A Alignment Panelists

Professional Position	Number of Panelists	Missouri Out-								
		1- SE	2- Heart	3- KC	4- NE	5- NW	6- SC	7- SW	8- STL	9- Central
Science: Grades 5, 8, 11										
Teacher	5	1	1				1	1	2	
Administrator Curriculum Specialist	1			1						

Materials

Panelists evaluated the alignment of the MAP-A performance tasks with the Missouri Show-Me Standards and AGLEs using forms for both the Webb and LAL alignment methods. A description of the forms, rating scales, and operational definitions is provided in Appendix B. All ratings were then coded into a Microsoft Excel spreadsheet to facilitate analyses.

Test Forms. Reviewers evaluated a sample of 30 Science MAP-A portfolios from the spring 2008 collection periods per grade level. Figure 1.1 from the previous chapter describes the content and structure of the portfolios. Briefly, each portfolio included two entries. Each entry was designed to measure one of six content stands and one of two process strands. So, each student’s portfolio was tailored to measure four AGLEs.

Rating Forms and Instructions. Panelists completed three rating forms individually and an additional three rating forms via group consensus (see Appendix B for descriptions of each form). Panelists received instruction sheets enumerating the alignment tasks that they needed to complete as well as code sheets listing the depth-of-knowledge ratings and other possible ratings for each task (see Appendix B).

Procedures

HumRRO conducted this alignment review in cooperation with the Missouri Department of Elementary and Secondary Education on July 13-15, 2009. Workshops were held at the Assessment Resource Center at the University of Missouri, Columbia. 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 then gave a brief presentation to describe alignment studies and to introduce tasks the reviewers would complete.

Following the general introduction, panelists began working within their content groups. A single group of seven panelists reviewed all Science standards documents and assessment materials. Other groups at the workshop reviewed Mathematics and Communication Arts materials.

Within the small group, a HumRRO staff member further trained reviewers using sample standards and assessment tasks. Regarding instructions on how to rate standards and items, the HumRRO staff member provided general suggestions and comments when appropriate; however, they emphasized to reviewers that staff would not provide explicit direction on how to rate standards or items because reviewers were valued as content experts. The HumRRO staff member provided brief instructions about how to use each rating form.

After reviewing sample DOK evaluations as a group, reviewers rated the Standards from the Missouri Show-Me Standards relevant to each grade-level test. Panelists first made independent evaluations without discussion. Once all reviewers had completed their ratings, groups discussed their ratings to achieve consensus DOK ratings for each Standard; a voluntary scribe within each group recorded these consensus ratings. Next, reviewers followed the same process to rate the DOK of the AGLEs, first individually and then to reach consensus.

Next, reviewers rated the AGLEs on a variety of factors, including (a) whether the Standard listed is the best match, (b) how well the AGLE links to the Standard, (c) whether the AGLE measures student performance of the Standard, (d) whether the AGLE is appropriate for the chronological age at which it is measured, (e) the level of symbolic communication required of students to demonstrate its content, and (f) whether the content expectation of the AGLE is accessible to various disability groups. These ratings were made individually; no consensus ratings were obtained.

Reviewers then received more specific instructions for rating portfolio performance tasks. For training, HumRRO staff facilitated reviewers in evaluating and discussing sample items as a group. After completing sample items, reviewers individually rated performance tasks on rating forms. The panelists rated the items on the same dimensions that they rated each AGLE (as described above). Reviewers in LAL alignment studies are typically instructed to assign a *primary Standard* to an item based on a judgment that an item clearly measured this Standard. Furthermore, reviewers could assign an *additional Standard* only in cases when the item seemed to assess another Standard as clearly as the primary Standard. For the Science MAP-A, the standard (AGLE) was clearly indicated by the teacher on each task. Teachers justified the event as a measure of a particular AGLE as part of the portfolio documentation. Therefore, reviewers verified that the teachers' indications of the standards the entries assessed were accurate and appropriate rather than matched the entry to standard themselves. Because this approach was confirmatory, results among reviewers indicated near exact agreement of which standard each portfolio entry assessed. Reviewers also indicated whether the content of the performance task

was academic and whether it could be modified or supports be provided without changing its meaning.

Finally, panelists worked in their small groups to develop consensus ratings for three additional aspects of the MAP-A Science assessment. HumRRO staff trained panelists on each task, and then the voluntary scribe from within the small group recorded the group's consensus ratings on rating sheets. The first consensus task required panelists to rate whole test barriers, or aspects of the Science MAP-A as a whole that might prevent students with various disabilities from fully participating (with or without supports or accommodations). The second consensus task asked panelists to rate the extent to which the scoring rubric and achievement standards allow for the demonstration of student learning. Lastly, reviewers developed consensus ratings of the extent to which content differs across grades.

Chapter 3 Results for MAP-A Science

Alignment of Science AGLEs to Missouri Show-Me Standards

LAL Criterion 1: Academic - *The content is academic and includes the major domains/strands of the content area as reflected in state and national standards (e.g., Mathematics, Communication Arts, Science).*

Per the USDE (2005), alternate assessments counting towards Title I must assess students only on academic content, as opposed to functional life skills. Panelists judged the grade-level Science assessments as to whether each AGLE focuses primarily on academics (similar to the Missouri Show-Me Standards). Results of this analysis are presented in Table 3.1. In terms of acceptability, at least 90% of AGLEs should be rated as academic.

Table 3.1. Mean Number of Science AGLEs Rated as Academic by Panelists

Grades	Number of AGLEs	Mean Number of AGLEs Academic		Mean Number of AGLEs Functional		Mean Percentage of AGLEs Rated Academic ^a	Number of Panelists Rating More than 90% of Rated AGLEs Academic ^a
		M	SD	M	SD		
5	64	64.0	0.0	0.0	0.0	100%	7 of 7
8	93	90.7	6.0	0.0	0.0	100%	7 of 7
11	211	211.0	0.0	0.0	0.0	100%	7 of 7

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

As Table 3.1 demonstrates, panelists indicated that the Science AGLEs focus primarily on academic content. Every panelist rated all of the AGLEs for all three grades as academic. The variability in the average *number* of AGLEs rated as academic for Grade 8 occurred because one rater neglected to provide ratings for a subset of the AGLEs. None of the AGLEs for any grade was rated as measuring functional skills, and all panelists met the criterion of rating more than 90% of the AGLEs as academic. Clearly, there was little debate among panelists about the academic nature of the Missouri AGLEs.

LAL Criterion 2: Age Appropriate - *The content is referenced to the student's assigned grade level (based on chronological age).*

This criterion pertains to the developmental level of the content included in the AGLEs. For this evaluation, panelists were asked whether the content of the Science AGLEs is appropriate for the age and grade level indicated. Response options for this scale included:

- Adapted - Linked to grade level content.
- Inappropriate - Content is off-grade level.
- Neutral - Content is not age-bound and is appropriate at any age.

Table 3.2 includes the results of panelists' ratings. Column 3 lists the rating categories, while the Mean in Column 4 refers to the mean number of AGLEs receiving that rating across panelists. Column 6 represents this same mean as a percentage of the total number of AGLEs per grade. Acceptability for this criterion is that at least 90% of AGLEs are rated as adapted or neutral⁸.

Table 3.2. Science AGLEs at Various Levels of Age Appropriateness

Grade	Number of AGLEs	Age Appropriateness Rating	Mean	SD	Mean Percentage of AGLEs per Rating ^a	Number of Panelists Rating at Least 90% of Rated AGLEs Adapted or Neutral ^a
5	64	Adapted	61.3	4.9	95.8%	6 of 7
		Neutral	0.1	0.4	0.2%	
		Inappropriate	2.6	4.5	4.0%	
8	93	Adapted	88.4	8.1	95.2%	6 of 7
		Neutral	0.0	0.0	0.0%	
		Inappropriate	4.4	8.2	4.8%	
11	211	Adapted	120.7	57.7	57.3%	0 of 7
		Neutral	7.7	13.8	3.7%	
		Inappropriate	82.3	50.3	39.1%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

As Table 3.2 displays, Grades 5 and 8 met the minimum criterion for age appropriateness, as panelists indicated on average that more than 90% of the AGLEs were adapted from grade-level content. However, at Grade 11, nearly 40% of the AGLEs were rated as inappropriate or off-grade. None of the panelists rated 90% or more of AGLEs as age appropriate at Grade 11, whereas six of the seven raters indicated at least 90% of the AGLEs for the elementary and middle school grades were age appropriate. Generally, fewer AGLEs met panelists' criteria for being age-appropriate at the upper grade, although results were more positive for Grades 5 and 8.

LAL Criterion 3: Standards Fidelity

- Content Centrality** - *The focus of achievement maintains fidelity with the content of the original grade level standards.*

⁸ The LAL method does not specify a minimum for Criterion 2. This minimum level was established by HumRRO.

To evaluate this criterion, panelists provided ratings indicating their judgments on the degree of content match between the AGLEs and Missouri Show-Me Standards for Science. First, we asked panelists to provide a simple evaluation (yes or no) of whether the Show-Me Standard listed as linked with the AGLEs did, in fact, match. For those AGLEs judged as matched to the designated standard, we then asked panelists to provide a second rating to indicate *how well* the AGLE linked to the standard.

Concerning overall content match, panelists at each grade level rated all (100%) of the Science AGLEs as matched to the Show-Me Standards.

For the second evaluation, panelists reviewed each grade-level AGLE for the degree of link to the central content targeted by the standards. In this case, panelists used the following 4-point scale to determine how well the AGLE reflects the standard content:

1	2	3	4
No Link	Weak Link	Moderate Link	Close Link

In terms of an acceptable level for this criterion, at least 90% of extended standards should be rated as 'moderate' or 'close' to the full standards. Table 3.3 shows that each set of grade-level AGLEs surpassed this minimum. Panelists found the majority of AGLEs to link sufficiently ('moderate' or 'close' link) with the standards, and no content AGLEs were rated as entirely different from the standards. No raters at Grades 5 and 8, and only one rater at Grade 11, rated fewer than 90% of AGLEs as moderately or closely linked.

Table 3.3. Mean Number of AGLEs at Various Levels of Content Centrality

Grade	Number of AGLEs	Content Centrality Rating	Mean	SD	Percentage of AGLEs per Rating	Number of Panelists Rating at Least 90% of AGLEs Moderate or Close ^a
5	64	No link	0.0	NA	0.0%	7 of 7
		Weak link	0.4	1.1	0.7%	
		Moderate link	2.7	7.2	4.2%	
		Close link	60.9	8.3	95.1%	
8	93	No link	0.0	NA	0.0%	7 of 7
		Weak link	0.7	1.9	0.8%	
		Moderate link	3.1	8.3	3.4%	
		Close link	88.7	10.1	95.4%	
11	211	No link	0.0	NA	0.0%	6 of 7
		Weak link	3.4	9.1	1.6%	
		Moderate link	9.0	23.8	4.3%	
		Close link	198.6	32.9	94.1%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

b. Performance Centrality - The focus of achievement maintains fidelity with the specified performance.

We asked panelists to directly compare the performance expectations in the AGLEs with the full content standards. Panelists evaluated the language of each AGLE in terms of whether the expectations are the same, partly similar, or differ entirely from what is expected in the corresponding standards. For example, if the standard requires students to compare and contrast traits, and the AGLE asks students to group or categorize based on traits, these expectations are parallel. If a standard expects students to identify and explain while the AGLE asks students to identify only, these expectations are partly similar. When students are asked to distinguish between in the standard but the AGLE requires students to recognize, then the expectation for demonstrating knowledge is different. Table 3.4 shows the results of this comparison. For acceptability, at least 90% of the AGLEs should be rated as partly similar or the same when compared with the full content standards.

Table 3.4. Mean Number of AGLEs at Various Levels of Performance Centrality

Grade	Number of AGLEs	Content Centrality Rating	Mean	SD	Percentage of AGLEs per Rating	Number of Panelists Rating at Least 90% rated Similar or Same ^a
5	64	Differ Entirely	0.0	NA	0.0%	7 of 7
		Partly Similar	5.3	14.0	8.3%	
		Same	58.7	14.0	91.7%	
8	93	Differ Entirely	0.1	0.4	0.2%	7 of 7
		Partly Similar	11.7	24.3	12.7%	
		Same	80.7	24.7	87.2%	
11	211	Differ Entirely	0.0	NA	0.0%	7 of 7
		Partly Similar	16.7	44.2	7.9%	
		Same	194.1	44.2	92.1%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

The AGLEs for each grade surpassed the minimum level of acceptability for all raters across grades, with the majority of the content expectations rated as requiring the same or a similar type of performance as the standards. Table 3.4 indicates that no AGLEs in Grades 8 or 11 received a rating of “Differ entirely” and only one AGLE was rated partly similar to the corresponding standards for Grade 5.

LAL Criterion 5: Content Differentiation - There is some differentiation in content across grade levels or grade bands.

This criterion focuses on whether the content expectations change appropriately between grades (e.g., whether the AGLEs for Grade 8 are sufficiently differentiated from the AGLEs for Grades 5 and 11). For this reason, the evaluation of content differentiation involves a comparison *between* grade-level content expectations. Panelists rated the AGLEs between grades as to whether they evidenced broader, deeper, and newer knowledge, as well as if certain expectations represented prerequisite skills (see Appendix B for a more detailed explanation of the categories). Across these categories, panelists indicated whether the content differentiation of AGLEs between grades was clear (C), partial (P), limited (L), or None (N). These ratings were reached collaboratively among panelists to achieve consensus evaluations. According to the LAL method, content expectations should show evidence of at least partial differences in content between grades on the dimensions of Broader, Deeper, Prerequisite, and New. Table 3.5 presents these results.

Table 3.5. Consensus Ratings on Content Differentiation between Grades MAP-A AGLEs for Science

Criterion	5	8	11	Selected Notes from Panelists
Broader	C	C	C	More APIs added at higher grades, and more subpoints (e.g., a, b, c, d) are added to AGLEs at higher grades
Deeper	C	C	C	Students need to do higher-level skills at higher grades; AGLEs move up through hierarchy of skills (e.g., explore to compare)
Prerequisite	C	C	C	Hierarchy of skills shows building complexity across grade levels
New	C	C	C	New AGLEs added at higher grade levels (e.g., Matter and Energy adds 1.4, 1.5, 1.6 at grades 5 to 1.1 and 1.2 which are at K-2)
Identical ^a	L	P	P	More AGLEs at the higher grades are identical with lower grades, and some AGLEs just have words added to those for lower grades

^a None (N) is an appropriate rating for this dimension because it indicates that no identical content is evident between grades.

As Table 3.5 demonstrates, panelists' ratings suggest an acceptable amount of content differentiation in Science AGLEs across grades. Generally, panelists indicated that AGLEs showed a clear progression in breadth, depth, and cognitive demands across grades. The organization of the AGLEs clearly enabled the panelists to observe the addition of new content at higher grades. One limited area of concern is the overlap among AGLEs at the upper grades; more of the material for these AGLEs is identical to content at lower grades. Overall, however, panelists indicated a good amount of content differentiation in the Science AGLEs.

LAL Criterion 7: Performance Accuracy - The potential barriers to demonstrating what students know and can do are minimized in the assessment to increase measurement accuracy of student performance.

Panelists evaluated whether students could reasonably demonstrate the content and performance expected in the AGLEs by providing several different ratings. First, we asked panelists to determine the level of communication required by each AGLE in order for students to demonstrate knowledge. The common categories applied, according to the LAL method, include the following three ability levels for students with significant disabilities⁹:

⁹ In addition to rating descriptions in the LAL manual, these definitions for communication levels have been expanded for clarity based on descriptions in a document published by the North Carolina Department of Public Instruction, Exceptional Children Division:
<http://www.ncpublicschools.org/docs/ec/instructional/extended/extendedcontentstandards.ppt>

- Pre-symbolic - student may demonstrate intentionality by showing interest, focus, or desire for a result through behavior; can use idiosyncratic gestures, sounds, or purposeful movements but no discrimination between pictures or other symbols.
- Early symbolic - student demonstrates emerging knowledge of symbols with some recognition of symbol-object relationships.
- Symbolic - student has broad knowledge of and can communicate consistently with symbols (e.g., pictures) or words (e.g., speech, assistive technology, signs).

In general for extended standards and alternate assessments, it is expected that teachers and test administrators modify the content to instruct and assess students at the appropriate level based on their IEPs. However, if the level of communication required in the extended standards document is always at the highest level (symbolic), it becomes more difficult for accommodations and supports to be provided to students at the more basic levels of communication and still retain comparability in content and performance. Teachers and assessment administrators find it much less problematic to increase the scope of content and performance expected for individual students than attempt to pare down. For these reasons, HumRRO's position on this issue is that it is preferable that the access point of most extended standards (and assessment tasks) be pre-symbolic¹⁰. Thus, the minimum level of acceptability is that the access point for at least 90% of the AGLEs should be pre-symbolic.

Table 3.6 presents panelists' mean ratings on the communication levels needed to demonstrate content knowledge for each set of AGLEs.

¹⁰ The authors of the LAL method suggest a different perspective that focuses more on symbolic communication. For more information, please refer to the *Links for Academic Learning: An Alignment Protocol for Alternate Assessments Based on Alternate Achievement Standards*.

Table 3.6. Mean Number of AGLEs Rated at Each Level of Symbolic Communication

Grade	Number of AGLEs	Level of Symbolic Communication Required	Mean	SD	Mean Percentage of AGLEs per Rating ^a	Number of Panelists Rating at Least 90% of Rated AGLEs at Pre-symbolic ^a
5	64	Pre-symbolic	39.8	12.5	62.2%	0 of 6 ^b
		Early symbolic	19.5	10.4	30.5%	
		Full symbolic	4.7	7.2	7.3%	
8	93	Pre-symbolic	33.7	25.7	36.2%	1 of 6
		Early symbolic	37.3	18.6	40.1%	
		Full symbolic	22.0	15.3	23.7%	
11	211	Pre-symbolic	75.7	38.1	35.9%	0 of 6
		Early symbolic	75.0	48.0	35.5%	
		Full symbolic	60.3	45.2	28.6%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

^b Data from only 6 raters are used in this analysis; one rater asked to have her data removed for this criterion because she realized she misunderstood the directions and consistently misrated her AGLEs.

Based on these panelists' ratings, none of the grades met the minimum acceptability of 90%. The highest percentage of Science AGLEs was at Grade 5, where, on average, panelists rated about 60% of the Science AGLEs as being accessible at the pre-symbolic level. Ratings for Grades 8 and 11 placed just a little over a third of the Science AGLEs at those grade levels at the pre-symbolic level; AGLEs were more similarly distributed across pre-symbolic, early symbolic, and full symbolic. These outcomes may indicate that students with the lowest level of symbolic abilities cannot access the full range of content expectations at any grade level, and we encourage Missouri to review the AGLEs to evaluate accessibility.

The second type of rating performed by panelists focused on general accessibility to students based on various types of disabilities (beyond communication abilities). For example, can students with visual impairments, an inability to follow instructions, or a need for assistive technology demonstrate the knowledge expected by these AGLEs? Panelists provided simple yes' (accessible to all) or no' (not accessible to some groups) responses to indicate their judgments. If they gave a no' rating, we asked panelists to provide some explanation of which groups would be disadvantaged and why in a Comments section. Table 3.7 includes the percentage of AGLEs that were judged as accessible to all groups.

Table 3.7. Science AGLEs Rated as Accessible to All Students

Grade	Number of AGLEs	Mean	SD	Mean Percentage of AGLEs Rated Accessible ^a	Number of Panelists Rating at Least 90% of AGLEs Accessible
5	64	63.9	0.4	99.8%	7 of 7
8	93	92.7	0.8	100%	7 of 7
11	211	208.0	6.7	98.6%	7 of 7

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

In contrast to their ratings on symbolic communication abilities, panelists felt that the Science AGLEs across all grades could be accessed by a wide range of students with different physical and cognitive disabilities. Nearly all Science AGLEs at all grades were rated as accessible to most students.

The third type of rating performed by panelists focused on the general extent to which accommodations or supports could be provided to enable students with various types of disabilities to access the content. For example, can students with hearing impairments or a need for assistive technology demonstrate the knowledge expected by these AGLEs if appropriate supports were provided? Panelists provided simple ‘yes’ (accessible with accommodations to all) or ‘no’ (not amenable to accommodations or supports for some groups) responses to indicate their judgments. If they gave a ‘no’ rating, we asked panelists to provide some explanation of which groups would be disadvantaged and why in a Comments section. Table 3.8 includes the percentage of AGLEs that were judged as accessible to all groups.

Table 3.8. Science AGLEs Rated as Amenable to Accommodations for All Students

Grade Span	Number of AGLEs	Mean	SD	Mean Percentage of AGLEs Rated Amenable ^a	Number of Panelists Rating at Least 90% of AGLEs Amenable
5	64	63.9	0.4	99.8%	7 of 7
8	93	91.7	3.4	100%	7 of 7
11	211	208.0	6.7	98.6%	7 of 7

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

As with the accessibility ratings, all panelists felt that the Science AGLEs across all grades were amenable to accommodations or supports that would enable them to be accessed by a wide range of students with different physical and cognitive disabilities. Nearly all Science AGLEs at all grades were rated as amenable to accommodations for most students. Because Missouri teachers select AGLEs to build tasks for their own students, they might be particularly aware of ways to accommodate the AGLEs to enable access for a wide variety of students.

Following their ratings of individual AGLEs, panelists were asked to reflect holistically on the entire set of AGLEs and reach consensus on the extent to which they include potential barriers that might limit student learning. Generally, panelists indicated that the Science AGLEs were widely accessible to most students. Because some of the AGLEs ask students to “explore” with their five senses, panelists indicated they were sufficiently accessible to students at all levels of communication; although the panelists only indicated about one third to two thirds of AGLEs were at the pre-symbolic level, they seemed to feel that the AGLEs as a whole offered ample opportunities for students at all levels. The only accessibility barriers that panelists pointed out were associated with some AGLEs for visually impaired students; panelists thought it would be difficult for these students to access AGLEs with clear visual elements (e.g., light and shadows, clouds).

Finally, we asked panelists to evaluate whether accommodations, modifications, and supports were sufficiently defined to enable standardized administration. Panelists acknowledged that the administration manual is specific about supports because they impact independence. However, accommodations and modifications are teacher based; individual teachers choose and apply AGLEs and any necessary accommodations for each specific student.

Alignment of Science Portfolio Assessments to AGLEs

LAL Criterion 1: Academic - *The content is academic and includes the major domains/strands of the content area as reflected in state and national standards (e.g., reading, mathematics, science).*

Per the USDE (2005), alternate assessments counting towards Title I must assess students only on academic content, as opposed to functional life skills. Panelists were asked to judge the Science assessments as to whether each task focuses primarily on academics. Results of this analysis are presented in Table 3.9. To be considered acceptable, at least 90% of tasks should be rated as academic.

Table 3.9. Mean Number of Science MAP-A Portfolio Tasks Rated as Academic by Panelists

Grades	Number of Tasks Rated	Mean Number of Tasks Academic		Mean Number of Tasks Functional		Mean Percentage of Tasks Rated Academic ^a	Number of Panelists Rating More than 90% of Rated AGLEs Academic ^a
		M	SD	M	SD		
5	30	29.3	1.1	0.7	1.1	97.6%	7 of 7
8	30	27.7	2.0	2.0	1.8	93.3%	6 of 7
11	30	28.6	1.7	1.4	1.7	95.2%	6 of 7

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

Generally, panelists rated the tasks administered on the MAP-A Science portfolios as academic in nature. At Grade 5, all 7 raters indicated that at least 90% of the tasks they rated were academic; in Grades 8 and 11, six of the seven raters indicated at least 90% of the tasks were academic. On the whole, more than 90% of the tasks were rated as academic across panelists at each of the three grades, and very few tasks were rated as functional. Thus, as with the Science AGLEs, panelists rated the MAP-A Science tasks as primarily academic in nature.

LAL Criterion 2: Age Appropriate - The content is referenced to the student's assigned grade level (based on chronological age).

Panelists evaluated the performance tasks on whether the content and performance assessed students at an appropriate level linked to their assigned grade. Table 3.10 shows the mean number and percentage of tasks judged as adapted (linked) to grade level, inappropriate (off-grade), and neutral (not age-bound). For acceptable linkage, at least 90% of tasks must be judged adapted or neutral. The Grade 5 Science MAP-A tasks surpassed the minimum requirement of 90% rated as “adapted” or “neutral”. At Grades 8 and 11, the 90% criterion was not quite achieved; 87% of tasks in Grade 8 and 84% of tasks for Grade 11 were rated as adapted or neutral. However, it is worth noting that, for both of these grades, one single rater rated all (Grade 11) or nearly all (Grade 8) of the tasks as inappropriate, pulling down the overall group average. Ultimately, most panelists found the majority of tasks to be linked to grade-level content; none of the other panelists rated more than 3 tasks as inappropriate at either grade level, and most raters indicated all tasks were appropriate.

Table 3.10. Science MAP-A Performance Tasks at Various Levels of Age Appropriateness

Grade	Number of Tasks Rated	Age Appropriateness Rating	Mean	SD	Mean Percentage of Tasks per Rating ^a	Number of Panelists Rating at Least 90% of Rated Tasks Adapted or Neutral ^a
5	30	Adapted	28.1	2.3	96.6%	7 of 7
		Neutral	0.3	0.8	1.0%	
		Inappropriate	0.7	1.1	2.5%	
8	30	Adapted	26.1	8.1	87.1%	6 of 7
		Neutral	0.3	0.8	1.0%	
		Inappropriate	3.6	8.2	11.9%	
11	30	Adapted	25.1	11.1	83.8%	6 of 7
		Neutral	0.1	0.4	0.5%	
		Inappropriate	4.7	11.2	15.7%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

LAL Criterion 3: Standards Fidelity

a. Content Centrality - *The focus of achievement maintains fidelity with the content of the original grade level standards.*

Panelists rated tasks for content match to the AGLEs to determine the extent to which the tasks assess grade-level content. Several analyses were performed on these ratings. First, panelists reviewed the number of tasks that were linked to at least one AGLE. The panelists for each grade felt that most of the tasks were aligned to the Science AGLEs as shown in Table 3.11. However, at each grade some tasks were rated to have no alignment. On average, more than 10% of the tasks were rated as not linked to the AGLEs for Grades 8 and 11. It is advisable to examine the tasks at these grades for relevance to the AGLEs.

Table 3.11. Mean Percent of Tasks Linked to AGLEs

Grade	Percentage of Tasks Linked to AGLEs
5	94.7%
8	83.8%
11	88.1%

We also asked panelists to evaluate *how well* the tasks targeted the AGLEs. For acceptability, at least 90% of tasks should be judged as moderately to closely linked with the AGLEs for acceptability. Table 3.12 presents the mean

number and percentage of tasks that fell into each category based on panelists' ratings.

Table 3.12. Mean Percent of Tasks at Various Levels of Content Centrality

Grade	Number of Tasks	Content Centrality Rating	Mean	SD	Percentage of AGLEs per Rating	Number of Panelists Rating at Least 90% of AGLEs Moderate or Close ^a
5	30	No link	1.0	0.0	3.4%	3 of 7
		Weak link	3.1	2.0	10.6%	
		Moderate link	10.4	4.3	35.1%	
		Close link	15.1	5.1	51.0%	
8	30	No link	2.0	1.6	6.8%	1 of 7
		Weak link	4.6	3.6	15.5%	
		Moderate link	8.6	3.6	29.0%	
		Close link	14.4	5.0	48.8%	
11	30	No link	2.7	1.9	9.1%	2 of 7
		Weak link	4.0	2.6	13.4%	
		Moderate link	9.6	4.9	32.1%	
		Close link	13.6	5.6	45.5%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

Panelists rated the majority of tasks across grades as linked sufficiently to the target content of the AGLEs. However, most panelists rated fewer than 90% as moderately or closely linked. These tasks may require review to determine if the content link could be improved with edits to the tasks. This is most problematic for Grades 8 and 11 where, on average, approximately 7 of 30 tasks were rated as having no link or a weak link.

b. Performance Centrality - The focus of achievement maintains fidelity with the specified performance.

In addition to the targeted content, the alternate assessment tasks should retain the performance intended by the full content standards to some extent. For example, if the full content standards require students to compare and contrast content, the AGLEs should require students to make some type of distinction. Table 3.13 shows the mean number of tasks rated as retaining all (same performance), some, or none of the performance expectations of the corresponding standards. (Would be clearer if descriptors in the table matched descriptors in this paragraph. i.e., some, none or all.) For acceptability, at least 90% of tasks should receive ratings of Some or All.

Table 3.13. Mean Percent of Tasks at Various Levels of Performance Centrality

Grade	Number of Tasks	Content Centrality Rating	Mean	SD	Percentage of AGLEs per Rating	Number of Panelists Rating at Least 90% of AGLEs Similar or Same ^a
5	30	Different	0.0	NA	0.0%	7 of 7
		Partly Similar	18.0	6.6	61.8%	
		Same	11.1	7.6	38.2%	
8	30	Different	3.1	1.5	10.5%	5 of 7
		Partly Similar	17.0	6.1	56.9%	
		Same	9.7	5.6	32.5%	
11	30	Different	1.7	1.1	5.8%	6 of 7
		Partly Similar	17.9	7.0	60.7%	
		Same	9.9	7.2	33.5%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

The test for Grade 5 met the minimum level of acceptability (90%) of tasks assessing students on at least some of the same performance expectations as the standards by all of the raters. Five raters at Grade 8 and 6 raters at Grade 11 felt that at least 90% of the tasks assessed the students on some of the same performance expectations as the standards.

LAL Criterion 4: Content Coverage (Webb dimensions) - The content differs from grade level in range, balance, and DOK, but matches high expectations set for students with significant cognitive disabilities.

For most alternate assessments, LAL Criterion 4 incorporates the Webb alignment statistics. These measures reveal the degree of content coverage, along with the extent of cognitive processing expected by the assessment. For example, one measure determines *which* content strands are covered by the assessment (referred to as Categorical Concurrence) based on panelists' judgments about the content targeted per test item or task. Results should correspond well with the state content standards and test blueprint documents. Panelists do not receive the test vendor's information about intended content target while giving their own ratings to retain an independent process.

The structure of the MAP-A as a portfolio-based assessment made the use of the Webb alignment statistics inappropriate. Many portfolio-based alternate assessments allow teachers to select the standard and corresponding entry (task) for their students. This information is included in the full portfolio as

part of the assessment, and frequently the scoring process incorporates the correspondence between the standard and task selected by teachers as part of the score. For the MAP-A alignment review, panelists received intact portfolios, which included identification of the targeted AGLEs along with the performance task description. Thus, no differences exist between reviewers regarding the AGLE matched to each task. Furthermore, it would have been difficult to impossible for reviewers to determine independently the AGLEs intended for assessment by teachers without some context for three reasons: (a) teachers provide explanation or rationale within the portfolio for how the task should relate to the AGLE, (b) the Missouri AGLE document includes a voluminous number of content standards, often overlapping, per strand at each grade, and (c) some tasks, unfortunately, were poorly written by teachers or were vague.

Despite differences in assessment structure, the goals of the Webb alignment indicators remain relevant; it is still important to determine whether teachers adequately cover the breadth and depth of the AGLEs. Although panelists had knowledge of the target content, they still had the ability to confirm or reject the appropriateness of teachers' selections. As a result, HumRRO addressed the following core alignment issues by examining the content and distribution of tasks relative to the AGLEs using simple frequency counts based on teachers' selections, as opposed to reviewers' ratings.

- How many content categories were covered by tasks (comparable to goal of *categorical concurrence* measure)?
- How many standards (AGLEs) within content categories were targeted (comparable to goal of *range-of-knowledge correspondence* measure)?
- Which standards per content categories were targeted most often by teachers? In other words, do teachers tend to distribute tasks across AGLEs, or do they write tasks for some AGLEs much more than others (comparable to goal of *balance-of-knowledge representation* measure)?
- Does the DOK of the entry match the DOK of the standard (true Webb measure of *depth-of-knowledge consistency*)?

Results for the Webb method are reported at the content strand level. The frequencies reported in tables below indicate the number of tasks that target each Science Strand based on teachers' selections identified in the portfolios. *If reviewers disagreed with teachers on the content targeted by tasks, we point to these discrepancies in the discussion below the reported results.*

Breadth Across Content Categories (similar to Categorical Concurrence). In the previous section on Content Centrality under LAL Criterion 3, we presented results on whether, and how well, each task matched to content expectations. For this analysis, we report on *which* AGLEs were assessed by teachers. As a minimum criterion to reflect adequate coverage per strand, we considered the assessment to reflect adequate breadth if at least half (50%) of

the total AGLEs within a strand were represented at least once in the sample of portfolios. This minimum criterion is less stringent than that established by Webb as well as the typical LAL alignment evaluation. We established a new criterion as Missouri's system would not meet the requirements of either because of the large numbers of AGLEs per strand and the small number of portfolio entries per subject area.

Table 3.14 summarizes the MAP-A alignment results for breadth across content categories. As Table 3.14 indicates, certain strands are assessed at each grade. Shaded areas represent content not assessed at that grade level. The process strands, scientific inquiry and science and technology, are assessed at all grade levels. Each portfolio entry is expected to address one of the content strands and one of the process strands—so each entry was coded for two standards. The '% accurate' column represents the panelists' agreement that the content indicated by teachers assessed the content of the standards. The final column indicates the number of standards represented within each strand in the sample of portfolios compared to the total number of standards included in that strand. For example, among the 15 portfolio entries coded for Matter and Energy, 6 of the 12 standards within that strand were represented. The strand for Scientific Inquiry (for all three grades) met our minimum criterion for breadth across content categories, as did Matter and Energy, Force and Motion, and Science and Technology for Grade 8 only.

It should be noted that this minimum criterion was established under the assumption that students would be instructed across multiple years and that the portfolio entries would be distributed across AGLEs representing various students instructional programs during the tested grade and prior grades. The criterion we established does not refer to individual student portfolios for a given year. If it did, the minimum criterion for breadth across content categories would not have been met for any strand at any grade. This new criterion was developed to provide Missouri with useful information about which strands are being assessed more completely than others. The criterion may be more similar to Webb's range-of-knowledge correspondence indicator than to categorical concurrence. Evidence from this table also is presented in the range-of-knowledge section.

Table 3.14. Summary of Breadth Across Content Categories Results for MAP-A Science Portfolios by Grade Level

Title of Strand	Number of Tasks per Strand							Standards with at Least One Task/Total Standards per Strand (Yes or No to Indicate if Minimum Criterion is Met)	
	Grade 5		Grade 8		Grade 11				
	Tasks Matched	% Accurate	Tasks Matched	% Accurate	Tasks Matched	% Accurate			
Matter and Energy			15	80.0			6/12 (Yes)		
Force and Motion			15	87.6			5/8 (Yes)		
Living Organisms	15	100					6/17 (No)		
Ecosystems	15	89.5					5/11 (No)		
Earth Systems					15	90.5	8/18 (No)		
Universe					15	85.6	5/21 (No)		
	Standards Assessed for All Grade Levels						Grade 5	Grade 8	Grade 11
Scientific Inquiry	16	92.0	15	80.0	15	84.8	2/3 (Yes)	3/3 (Yes)	3/4 (Yes)
Science, Technology	14	98.0	15	87.7	15	91.4	1/3 (No)	4/7 (Yes)	5/12 (No)

Depth-of-Knowledge Consistency. Depth-of-knowledge (DOK) consistency measures the type of cognitive processing required by each performance task compared to the requirements implied by the content objectives. In this case, ratings on the MAP-A portfolios can be analyzed using the common DOK consistency measure established by Webb because reviewers made these judgments independently for the tasks and the AGLEs. Teachers do not identify DOK levels per task.

As part of the rating process, reviewers first determined the DOK level for each AGLE using a rating scale (see Appendix B for the LAL DOK level descriptions). Next, as they reviewed performance tasks, panelists rated the level of processing needed to perform the task using the same DOK rating scales. We compared these separate ratings on cognitive complexity to determine the extent to which the assessed performance matched the content expectations specified in the AGLEs. Tables 3.15-3.17 summarize the DOK consistency results for each grade level of the Science MAP-A assessment. Since reviewers evaluated DOK at the most specific level of the standards document (AGLEs), the table refers to consistency between the tasks and the AGLEs to which they were matched. Results are summarized in terms of the percentage of AGLEs assessed by tasks at or above the same cognitive level. Webb’s suggested criterion for this alignment indicator is the same as for a regular assessment – that is, at least 50% of the tasks should have complexity ratings at or above the level of the corresponding AGLE per content strand. The minimum criterion for DOK consistency was met for all strands except Scientific Inquiry for Grades 5 and 8.

Table 3.15. DOK Consistency for Science MAP-A, Grade 5: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives

Title of Strand	Mean Tasks per Strand	DOK Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
Matter and Energy								
Force and Motion								
Living Organisms	15	0	0	29.5	0.51	70.5	0.57	Yes
Ecosystems	15	0	0	27.6	0.48	72.4	0.68	Yes
Earth Systems								
Universe								
Scientific Inquiry	16	52.7	0.75	35.7	0.62	11.6	0.65	No
Science, Technology	14	0	0	18.4	0	81.6	0.66	Yes
Percentage of strands with 50% of item DOK at or above objective DOK:								75%

Table 3.16. DOK Consistency for Science MAP-A, Grade 8: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives

Title of Strand	Mean Tasks per Strand	DOK Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
Matter and Energy	15	19.0	1.00	43.8	0.54	37.1	0.67	Yes
Force and Motion	15	28.6	0.94	31.4	0.48	40.0	0.55	Yes
Living Organisms								
Ecosystems								
Earth Systems								
Universe								
Scientific Inquiry	15	51.4	0.90	27.6	0.35	21.0	0.66	No
Science, Technology	15	7.6	0.55	11.4	0.51	81.0	0.65	Yes
Percentage of strands with 50% of item DOK at or above objective DOK:								75%

Table 3.17. DOK Consistency for Science MAP-A, Grade 11: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives

Title of Strand	Mean Tasks per Strand	DOK Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
Matter and Energy								
Force and Motion								
Living Organisms								
Ecosystems								
Earth Systems	15	28.6	0.77	21.0	0.78	50.5	0.55	Yes
Universe	15	6.7	0.98	30.5	0.80	62.9	0.75	Yes
Scientific Inquiry	15	32.4	0.98	52.4	0.13	15.2	0.68	Yes
Science, Technology	15	3.8	0.58	16.2	0.70	80	0.62	Yes
Percent of strands with 50% of item DOK at or above objective DOK:								100%

Table 3.18 shows how the DOK ratings were distributed for Science. Ratings were from 1—6 indicating attention, memorize/recall, performance, comprehension, application, and analysis/synthesis/evaluation, respectively. The mean, standard deviation, and range are presented to provide an indication that typical DOK level portfolio entries are designed to assess. As can be seen in the table, most entries were designed to assess at the memorize/recall and performance DOK levels. All but the highest DOK level (analysis/synthesis/evaluation) were represented at all grades.

Table 3.18. Mean Percentage of Tasks at Each DOK Level

Grade	Mean	SD	Range
5	2.62	0.90	1-5
8	2.65	1.00	1-5
11	2.80	0.90	1-5

Breadth within Content Categories (similar to Range-of-Knowledge Correspondence). Webb's range-of-knowledge measure indicates how fully the performance tasks cover each of the AGLEs within each major content category. The assessed AGLEs within a strand should be linked with at least one performance task. Webb's minimum level of acceptability for range-of-knowledge correspondence is that a mean of 50% of standards per content category should be matched to at least one assessment task. For the calculation of range for the MAP-A, we determined the frequency of performance tasks matched to each content strand by teachers. The minimum level of acceptability in this case was set at 50% of AGLEs per content strand matched to at least one task per grade across all portfolios in the sample. We used the same criterion to indicate breadth across content categories under the assumption that students would have multiple learning opportunities across a grade span (for example, the assessed Grade 5 and prior Grades 3 and 4) and that the portfolios represent a sampling of them. This assumption allows us to aggregate across portfolios. A conventional interpretation of Webb's breadth criterion is inappropriate because no single student's portfolio (with 2 entries coded for four AGLEs) could meet the minimum requirement. The final column in Table 3.14 above contains this information as well as a summary indication of whether this minimum criterion was met. The strand for Scientific Inquiry (for all three grades) met our minimum criterion for range (the same criterion used to indicate breadth across content categories above), as did Matter and Energy, Force and Motion, and Science and Technology for Grade 8 only.

Distribution of Assessment Tasks Among AGLEs (similar to Balance-of-Knowledge Representation). The fourth measure of alignment included in the Webb method is *balance-of-knowledge representation*. This measure indicates the number of tasks linked to each standard per strand. The number of tasks should be distributed rather evenly between the standards for each strand to achieve a balanced assessment. The content balance is determined by calculating an index, or score, for each strand¹¹. The large number of AGLEs per strand and the relatively small number of portfolio entries make this indicator inappropriate for the MAP-A. Each portfolio typically assesses four different strands, and consequently four different AGLEs. The balance indicator would indicate that the content is evenly represented across strands, but this indication would not have meaning in this context. Table 3.14 shows that among the 30 portfolios sampled per grade level, 15 typically represent each assessed strand. Tables in Appendix A show the number of entries per AGLE to give DESE an indication of which standards are being emphasized, but no index was calculated for this aspect of content coverage.

¹¹ The exact formula for calculating the balance index is explained in detail in Norman Webb's (2005) alignment training manual: <http://www.wcer.wisc.edu/WAT/index.aspx>.

Summary and Recommendations for Content Coverage. The MAP-A Science assessment is a portfolio system with only two entries per student. Each entry is designed to measure two AGLEs (one content and one process). Two science content strands are assessed within each grade. The process strands are assessed at all grade levels. There are as many as 21 AGLEs within a single strand and at least 8 AGLEs within each content strand. As such, typical measures of content coverage, such as Webb’s alignment criteria, are not appropriate; there is simply no way to represent the breadth of the content represented by so many AGLEs in a single portfolio. The MAP-A science assessment does not meet traditional minimum alignment criteria for breadth of content coverage, breadth within content categories, or distribution of assessment tasks among AGLEs.

In order to provide Missouri with information regarding the relative strengths and weaknesses of the science MAP-A, we aggregated portfolios under the assumption that students could receive instruction on multiple AGLEs within strands as they progressed from grade to grade, even though the assessment occurs only in Grades 5, 8, and 11. We also established new minimum criteria as a means of pointing out where the largest gaps in content coverage across portfolios occurred. These indicators are informative only, and should not be used as substitutions for more stringent alignment indicators.

The one exception to the rule for these analyses was DOK consistency. It was possible to directly compare the DOK of the portfolio entries with that indicated by the AGLEs. With the exception of Scientific Inquiry at Grades 5 and 8, the portfolios met Webb’s criteria of 50% or more entries at or above the DOK of the AGLE. Teachers seem to be preparing portfolio tasks for students who have DOK requirements that are as high as or higher than indicated by the standards.

There are currently too many AGLEs to be represented by any student’s portfolio. This would be true even if the portfolio was tripled in size (e.g. aggregated across 3 grade levels). Missouri may wish to consider either reducing the number of AGLEs to be assessed or altering the manner of assessment to include additional assessment items. There is currently no way to ensure that the assessment scores represent the scope of the Science content indicated by the standards.

LAL Criterion 5: Content Differentiation - There is some differentiation in content across grade levels or grade bands.

As with the evaluation of the AGLEs, LAL Criterion 5 focuses on whether the content increases in depth, breadth, and complexity at higher grade levels. Panelists achieved consensus ratings on the amount of content differentiation of the Science MAP-A performance tasks between grade levels (higher and lower). Table 3.19 shows panelists’ consensus ratings across the various dimensions using the rating scheme of clear (C), partial (P), limited (L), or none (N). For

acceptability, each test should demonstrate evidence that it possesses at least partially different content per dimension relative to higher or lower grade tests.

Table 3.19. Consensus Ratings on Content Differentiation between Grades MAP-A Assessments for Science

Criterion	5	8	11	Selected Notes from Panelists
Broader	N	N	L	Tasks look the same at grades 5 and 8; the same activities appeared again and again across grade levels. At grade 11, there is a little more evidence of differentiation. Hard to judge because different strands assessed at different grades.
Deeper	L	L	P	Mostly, tasks were written to the same level of complexity. There was very limited number of examples with deeper-level tasks from grade 5 to grade 8, but the complexity of tasks was higher at 11.
Prerequisite	L	L	N	For the most part, skills didn't seem to build on each other. Sometimes tasks might have been building blocks, but generally seemed similar. This is hard to judge because we are not seeing the same students over time and assessed strands differ.
New	L	L	P	Saw different tasks at higher grade levels, but largely because different content strands were assessed at different grades. Even despite that, many of the tasks were highly similar across grades. Some new skills/knowledge were introduced at 11.
Identical ^a	P	P	P	Most tasks were the same across grades, but there were a very few instances where things weren't identical. At grade 11, some new knowledge and skills were introduced.

^a None (N) is an appropriate rating for this dimension because it indicates that no identical content is evident between grades.

Panelists' ratings of content differentiation for the Science MAP-A performance tasks indicate a great deal of similarity across grades. Although panelists had observed clear differentiation of Science AGLEs across grades, the differentiation in content expectations did not carry over into the actual performance tasks administered on the MAP-A. The lack of differentiation was more pronounced at Grades 5 and 8; panelists indicated a moderate increase differentiation at Grade 11, where some additional skills or knowledge were sometimes required for performance tasks. One inherent difficulty in conducting these ratings specific to Science is that different content strands are assessed at different grade levels. Thus, coverage of different material in different grade levels can be more reflective of the requirements for assessment than a true progression in material over time. Even despite the different content strands to be assessed at different grades, panelists often observed highly similar performance tasks over time, particularly in the elementary and middle school grades.

LAL Criterion 6: Achievement - *The expected achievement for students is for the students to show learning of grade-referenced academic content.*

The sixth LAL criterion pertains to demonstration of student learning. The focus in this case is whether students are given the opportunity to demonstrate academic skills or knowledge acquired from their coursework on the assessment. To determine the extent to which the MAP-A *enables* students to demonstrate this learning, panelists evaluated the scoring rubric and achievement level descriptors relative to the assessment. Panelists worked together to determine whether the assessment allowed for demonstration of high, low, or no evidence of student learning. These consensus ratings were made across several dimensions of learning, which are described below (adapted from Flowers et al, 2007):

- Level of accuracy - extent to which scoring makes clear distinctions in student responses.
- Level of independence - extent to which student performance is based on independent response without teacher supports.
- New learning - extent to which evidence of new learning is demonstrable based on use of baseline or pretest OR clear content differentiation between grade tests.
- Generalization across people and settings - extent to which students must demonstrate knowledge across people or settings to receive credit.
- Generalization across materials and activities - extent to which students must demonstrate knowledge across different types of materials (i.e., objects) or activities.
- Standard setting - extent to which achievement standards are distinct and based on demonstration of independent student performance.
- Program quality indicators - extent to which the inclusion of program characteristics (i.e., opportunities for instruction; access to materials; teacher qualities) is limited as part of student score.

For accurate assessment of achievement, most dimensions should receive ratings of high inference regarding the ability to evaluate student learning.

Table 3.20 includes the group consensus ratings on the degree of student inference evident in the Science MAP-A assessment per grade level.

Table 3.20. Degree of Inference Evident on Student Learning in Science MAP-A Assessments

Dimensions	Grade 5	Grade 8	Grade 11	Selected Notes from Panelists
Level of Accuracy	H ^a	H	H	MO has standards for accuracy, but teachers almost always pick tasks where students can attain, so not much variability. Almost all portfolios have high levels.
Level of Independence	H	H	H	MO has guidelines for level of independence in manual. No hand-over-hand tasks observed in portfolios. Prompts/guidance seemed non-task-related (e.g. focus, listen, re-direct).
New Learning	N ^c	N	N	This is difficult to rate because different strands are assessed at different grades. Multiple student work records often involve different tasks. Same tasks repeated across years, and same APIs or tasks can be used although these need to be justified. This is not a progress-based test.
Generalization Across People and Settings	H	H	H	Because tasks have an application piece, almost every task had 2 pieces. For instance, asked to do something in the classroom and then in the grocery store (students actually physically go). Tasks demand real-life applications. Also make connections across content areas. These connections increase at upper grade levels, where students conduct tasks in even more settings.
Generalization Across Materials and Activities	L ^b	L	L	Most APIs only have one task. Lots of tasks used water or rocks.
Standard Setting	L	L	L	Achievement-level descriptors aren't measurable—what is 'strong' vs. 'weak'? The ALDs do specify level of independence necessary for credit. Terms aren't well defined, leave room for interpretation.
Program Quality Indicators	L	L	L	Whole score is based on application to standards, accuracy and independence. No impact of PQIs. However, teacher must choose and justify API; if teacher makes an error, it does lower portfolio score.

^a H = high student inference

^b L = low student inference

^c N = no student inference

As presented in Table 3.20, panelists' ratings of the Achievement criterion for the Science MAP-A were mixed. According to panelists' consensus ratings, the Science MAP-A enables a high degree of inference about student learning across all grades on the dimensions of Level of Accuracy, Level of Independence, and Generalization Across People and Settings. Generally, panelists' comments indicated that students needed to perform the MAP-A tasks

independently and that any prompts from the teacher tended to be non-task-related (e.g., re-directing attention). Generalization Across People and Settings was rated as a clear strength of this assessment; because each task requires two AGLEs and a real-life application of the skill being measured, generalization is an integral aspect of the test. Panelists indicated that the Science MAP-A enables limited inferences of student learning on the dimensions of Generalization across Materials and Activities, Standard Setting, and Program Quality Indicators. They noted that most AGLEs are represented with only one task and that many tasks involved the same materials, such as water or rocks. Regarding the MAP-A Science achievement level descriptions, panelists suggested it would be valuable to develop more measureable definitions to aid in interpretation. Finally, the dependence of students' scores on teacher care and effort led panelists to assign a rating of L for Program Quality Indicators because teachers' work can impact students' performance. Finally, panelists indicated no inferences about student learning could be made on the New Learning dimension because students might take the same task or AGLE across the grades; panelists struggled with this rating because the Science MAP-A assesses different content strands at different grade levels, which makes it more difficult to consider the role of prerequisite skills.

LAL Criterion 7: Performance Accuracy - *The potential barriers to demonstrating what students know and can do are minimized in the assessment to increase measurement accuracy of student performance.*

This criterion is intended to evaluate the degree of accessibility of the assessment for all student groups who take it. Reduced access to the assessment tasks would decrease accurate measurement of these students' skills. As with the AGLEs, panelists rated tasks on the levels of communication required to respond and the access available to each type of student who takes the assessment. In addition, panelists evaluated each task on whether accommodations or supports can be provided for different types of students without substantially altering the target content.

Table 3.21 gives mean ratings by reviewers on the communication levels required of students in order to respond to the Science tasks. For acceptability, at least 90% of tasks should be rated as pre-symbolic for reasonable access by all students.

Table 3.21. Mean Number of Tasks Rated at Each Level of Symbolic Communication

Grade	Number of Tasks Rated	Level of Symbolic Communication Required	Mean	SD	Mean Percentage of Tasks per Rating ^a	Number of Panelists Rating at Least 90% of Rated Tasks at Pre-symbolic ^a
5	30	Pre-symbolic	8.4	5.8	28.1%	0 of 7
		Early Symbolic	16.7	5.5	55.7%	
		Full Symbolic	4.9	1.2	16.2%	
8	30	Pre-symbolic	10.9	8.8	36.2%	0 of 7
		Early Symbolic	12.1	6.6	40.5%	
		Full Symbolic	7.0	4.8	23.3%	
11	30	Pre-symbolic	5.9	6.5	19.5%	0 of 7
		Early Symbolic	15.7	7.2	52.4%	
		Full Symbolic	8.4	7.7	28.1%	

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

Based on these panelists' ratings, none of the grade levels met the minimum requirement of 90%. For all three grades, the largest percentage of performance tasks was rated at the early symbolic level. Although these ratings do not meet the minimum requirement of 90% at the pre-symbolic level, their interpretation is complicated by the fact that they are based on individually designed portfolios. Because the tasks being rated are designed for individual students, it is optimal for teachers to create tasks at a level of communication appropriate for each student. Thus, if most of the students who performed the portfolio tasks rated for this study were functioning at an early symbolic level, the tasks may very well have been appropriate for those students. These ratings at the task level have less meaning for portfolios because the tasks were not designed to be accessible to a maximal amount of students; they were designed to be accessible and appropriate for the particular student taking the assessment.

Because portfolio tasks are designed for individual students, panelists were not asked to provide ratings of accessibility for the Science performance tasks as they did for the AGLEs. However, panelists did rate the general extent to which tasks were amenable to appropriate supports or accommodations. Panelists provided simple 'yes' (amenable to accommodations or supports) or 'no' (not amenable to accommodations or supports) responses to indicate their judgments. If they gave a 'no' rating, we asked panelists to explain their rationale in a Comments section. Table 3.22 includes the percentage of AGLEs that were judged as amenable to accommodations.

Table 3.22. MAP-A Science Performance Tasks Rated as Amenable to Accommodations for All Students

Grade	Number of Tasks Rated	Mean	SD	Mean Percentage of Tasks Rated Amenable ^a	Number of Panelists Rating at Least 90% of AGLEs Amenable
5	30	30.0	0.0	100.0%	7 of 7
8	30	29.3	1.0	97.6%	7 of 7
11	30	29.6	0.8	98.6%	7 of 7

^a Percentages are based on actual ratings. Missing data were excluded from the numerator and denominator.

As with the Science AGLEs, all panelists felt that the MAP-A Science performance tasks across all grades were amenable to accommodations or supports that enabled them to be accessed by a wide range of students with different physical and cognitive disabilities. Again, these high ratings might be a function of the fact that MAP-A tasks are designed by considering particular student needs.

Following their individual ratings of the Science MAP-A performance tasks, panelists were asked to reflect holistically on the entire set of tasks and to reach consensus on the extent to which they include potential barriers that might limit student learning. Indeed, panelists indicated that the specialized nature of students' performance tasks removed all barriers to students' ability to demonstrate their learning. As with the AGLE ratings, panelists acknowledged a lack of standardization across tasks; they described accommodations and modifications as —almost =built-in” because of the extent to which they are tailored to individual students. Assessment of students with disabilities often requires trade-offs between accessibility and standardization, and it appears panelists indicated that the MAP-A favors accessibility and opportunities for students to demonstrate learning over standardization in administration across students.

Inter-Rater Agreement Results

We evaluated the extent to which panelists provided exactly the same ratings on items, which qualifies as a measure of absolute agreement (Shavelson & Webb, N. M., 2005; Tinsley & Weiss, 1975). Most of the LAL criteria require categorical ratings (e.g., Which AGLE is appropriate? Is AGLE accessible or not?) on the AGLEs and portfolio tasks. Several agreement measures exist to analyze categorical ratings (see Gwet, 2001; Webb, N. L., 2005). For these data, we applied a measure based on one developed by Norman Webb, which provides a basic estimate of percent agreement between reviewers¹². This analysis involves a pair-wise comparison (one-to-one) of each reviewer's ratings

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.

with all other reviewers per item or task. Results then are averaged across reviewers per test form and evaluated as follows:

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

As can be seen in Table 3.23, panelists agreed substantially in their ratings with good to exact agreement on six of the eight dimensions. The additional dimensions (API match and Communication Levels) were at the higher end of the adequate reliability range.

Table 3.23. Pairwise Comparisons on AGLE Ratings per Grade Level

LAL Criteria	Min Agree	Max Agree	Mean across Strands	SD
Academic	86%	100%	99%	3%
API Match	29%	100%	67%	20%
Content Centrality	71%	100%	95%	7%
Performance Centrality	71%	100%	91%	8%
Age Appropriate	57%	100%	84%	15%
Communication Levels	43%	100%	69%	16%
Accessibility	71%	100%	99%	4%
Accoms_Supps	71%	100%	99%	4%

Table 3.24 presents these same types of agreement analyses on panelists' ratings of the portfolio tasks. Again, an exact match indicates that all reviewers agreed in their ratings across the Strand level. The panelists had good to excellent average levels of agreement on four of the portfolio dimensions, and adequate agreement on the remaining three dimensions.

Table 3.24. Pairwise Comparisons on Portfolio Ratings per Grade Level

LAL Criteria	Min Agree	Max Agree	Mean across Strands	SD
Academic	57%	100%	93%	11%
API Match	57%	100%	91%	12%
Content Centrality	29%	100%	61%	17%
Performance Centrality	43%	100%	65%	16%
Age Appropriate	57%	100%	84%	11%
Communication Levels	43%	100%	60%	14%
Accoms_Supps	57%	100%	95%	9%

References

- Flowers, C., Wakeman, S., Browder, D., & Karvonen, M. (2007). *Links for academic learning: An alignment protocol for alternate assessments based on alternate achievement standards*. Charlotte, NC: University of North Carolina at Charlotte. Retrieved from:
http://www.naacpartners.org/LAL/documents/NAAC_AlignmentManualVer8_3.pdf
- Missouri Department of Elementary and Secondary Education, Measured Progress, Assessment Resource Center. *Missouri Assessment Program-Alternate (MAP-A): Instructor's guider and implementation manual*. Roseville, MN: Missouri Department of Elementary and Secondary Education. Retrieved from:
<http://www.dese.mo.gov/divimprove/assess/documents/2009-2010-MAP-A-Web-Instructors-Guide.pdf>
- No Child Left Behind Act of 2001. Public Law 107-110.
- North Carolina Department of Public Instruction. (unknown). *Extended content standards: Three levels of access so that all children can participate in the general education curriculum*. Charlotte, NC: North Carolina Department of Public Instruction. Retrieved from:
<http://www.ncpublicschools.org/docs/ec/instructional/extended/extendedcontentstandards.ppt>
- Thompson, S.J., Johnstone, C.J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Missouri, National Center on Educational Outcomes.
- Thompson, S. J., Morse, A. B., Sharpe, M., & Hall, S. (2005). *Accommodations manual: How to select, administer, and evaluate use of accommodations for instruction and assessment of students with disabilities*. CCSSO State Collaborative on Assessment and Student Standards Assessing Special Education Students. Washington, DC. Retrieved from
http://www.osepideasthatwork.org/toolkit/accommodations_manual.asp
- U.S. Department of Education. (August, 2005). *Alternate achievement standards for students with the most significant cognitive disabilities*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education. Retrieved from
<http://www.ed.gov/admins/lead/account/saa.html#guidance>.

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

Appendix A
Alignment Results per Assessment

Alignment Results

The following tables include complete statistical results similar to the Webb alignment indicators (LAL Criterion 4: Content Coverage). There are several deviations from the typical Webb alignment method that were required due to the nature of the MAP-A portfolio entries. Deviations from the Webb method are explained by alignment indicator below.

Breadth Across Content Categories

The breadth across content categories (similar to Webb's Categorical Concurrence) results for Grades 5, 8, and 11 for the Science MAP-A are presented in Tables A-1-3 below. Each table includes: the title of the strand (broadest analyzed category); the number of portfolio entries matched to that strand; and the percentage of portfolio entries determined by panelists to accurately represent the targeted content. This table deviates from typical categorical concurrence reports for several reasons. First, the data included in the table does not represent the panelists matching content descriptions from the standards to portfolio entries. Each portfolio entry was marked with the standard it was designed to measure by the students' teachers. Panelists were asked to verify that the standards indicated by the teachers were accurate. There was therefore no variance among the panelists regarding the match to standard. Second, the number of portfolio entries per student was only 4. It would not make sense to expect each portfolio to represent all of the standards under each strand, so Webb's typical criterion of determining whether each strand is represented by at least one item is never achieved, nor reported. Similarly, because only 15 portfolios were analyzed per grade, it would not be expected that every standard be represented across all portfolios. Also, analyses across portfolios would not be indicative of the concurrence between the standards and the assessment for any student. For this table, we chose not to include an indication of whether Webb's categorical concurrence criterion was met.

Table A-1. Content Representation for Science MAP-A: Number of Portfolio Entries per Strand

Title of Strand	Number of Tasks per Strand					
	Grade 5		Grade 8		Grade 11	
	Tasks Matched	% Accurate	Tasks Matched	% Accurate	Tasks Matched	% Accurate
Matter and Energy			15	80.0		
Force and Motion			15	87.6		
Living Organisms	15	100				
Ecosystems	15	89.5				
Earth Systems					15	90.5
Universe					15	85.6
Scientific Inquiry	16	92.0	15	80.0	15	84.8
Science, Technology	14	98.0	15	87.7	15	91.4
Total	60		60		60	

Depth-of-Knowledge Consistency

The Depth-of-Knowledge (DOK) consistency results for Grades 5, 8, and 11 of the Science MAP-A assessment are presented below. The tables present the results from the comparisons between the depth of knowledge expected in the standards and the DOK assessed by items. The tables include the mean percentage of items rated as below, at the same level, or above the DOK level of the content standards along with the corresponding standard deviations. Results are separated by grade. Standards with at least 50% of items at the same (or above) DOK level met the minimum criterion. The percent of strands summary percentage in the last row of the tables refers only to strands assessed by the portfolios in the sample (does not include non-assessed strands).

Table A-2. Depth-of-Knowledge Consistency for Science MAP-A, Grade 5: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives

Title of Strand	Mean Tasks per Strand	DOK Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
Matter and Energy								
Force and Motion								
Living Organisms	15	0	0	29.5	0.51	70.5	0.57	Yes
Ecosystems	15	0	0	27.6	0.48	72.4	0.68	Yes
Earth Systems								
Universe								
Scientific Inquiry	16	52.7	0.75	35.7	0.62	11.6	0.65	No
Science, Technology	14	0	0	18.4	0	81.6	0.66	Yes
Percent of strands with 50% of item DOK at or above objective DOK:								75%

**Table A-3. Depth-of-Knowledge Consistency for Science MAP-A, Grade 8:
Mean Percent of Performance Tasks with DOK Below, At, and Above DOK
Level of Objectives**

Title of Strand	Mean Tasks per Strand	DOK Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
Matter and Energy	15	19.0	1.00	43.8	0.54	37.1	0.67	Yes
Force and Motion	15	28.6	0.94	31.4	0.48	40.0	0.55	Yes
Living Organisms								
Ecosystems								
Earth Systems								
Universe								
Scientific Inquiry	15	51.4	0.90	27.6	0.35	21.0	0.66	No
Science, Technology	15	7.6	0.55	11.4	0.51	81.0	0.65	Yes
Percent of strands with 50% of item DOK at or above objective DOK:								75%

Table A-4. Depth-of-Knowledge Consistency for Science MAP-A, Grade 11: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives

Title of Strand	Mean Tasks per Strand	DOK Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
Matter and Energy								
Force and Motion								
Living Organisms								
Ecosystems								
Earth Systems	15	28.6	0.77	21.0	0.78	50.5	0.55	Yes
Universe	15	6.7	0.98	30.5	0.80	62.9	0.75	Yes
Scientific Inquiry	15	32.4	0.98	52.4	0.13	15.2	0.68	Yes
Science, Technology	15	3.8	0.58	16.2	0.70	80	0.62	Yes
Percent of strands with 50% of item DOK at or above objective DOK:								100%

Distribution of Tasks Among AGLEs

Tables A-5-A-7 provide the distribution of AGLEs represented by portfolio entries. These numbers are aggregates and are provided to give DESE an indication of which AGLEs are emphasized and which might be avoided by teachers as they prepare the performance tasks. Typically the earlier AGLEs within a strand, especially for the process strands, are emphasized.

Table A-5. Summary of Portfolio Entry-to-AGLE Distribution Results for MAP-A Science for Grade 5

AGLE	Frequency of Tasks per Content strand			
	Strand 3	Strand 4	Strand 7	Strand 8
1	3	5	11	14
2	4	3	5	
3	2			
4		2		
5	3	3		
7		2		
11	2			
16	1			

Table A-6. Summary of Portfolio Entry-to-AGLE Distribution Results for MAP-A Science for Grade 8

AGLE	Frequency of Tasks per Content strand			
	Strand 1	Strand 2	Strand 7	Strand 8
1	6	6	8	6
2	1	3	5	5
3		2	2	3
4	1			1
5	3	3		
6		1		
7	1			
8	2			

Table A-7. Summary of Portfolio Entry-to-AGLE Distribution Results for MAP-A Science for Grade 11

AGLE	Frequency of Tasks per Content strand			
	Strand 5	Strand 6	Strand 7	Strand 8
1	4	3	6	6
2	1		8	2
3	1	9	1	2
4	3	1		3
5	3			2
6	1			
7		1		
9	1			
10		1		
12	1			

Appendix B
Workshop Instructions

MAP-A Panelist Task Instructions

	Rating Task	Documents Needed
1 a	DOK of Missouri Show-Me Standards (individual and consensus)	(1) Select Missouri Show-Me Standards (2) Rating Scale Code Descriptions – DOK scale
1 b	DOK of A-GLEs (individual and consensus)	(1) Alternate Grade-Level Expectations (A-GLEs) for your content area (2) Rating Scale Code Descriptions – DOK scale (3) A-GLE_DOK Rating Forms
2	Alignment Dimensions of A-GLEs (individual)	(1) Alternate Grade-Level Expectations (A-GLEs) for your content area (2) Rating Scale Code Descriptions (3) A-GLE_Alignment Dimensions Rating Forms (4) MAP-A Test Documents
3	Alignment Dimensions of Portfolio Entries (individual)	(1) Alternate Grade-Level Expectations (A-GLEs) for your content area (2) Rating Scale Code Descriptions (3) Portfolio_Alignment Dimensions Rating Forms (4) MAP-A Test Documents a. Instructor Administration Manual b. NCEO Considerations Guide c. Portfolios (accessible online)
4	Content differentiation across grades (consensus)	(1a) Alternate Grade-Level Expectations (A-GLEs) for your content area (1b) Rating Scale Code Descriptions – Content Differentiation (1c) Content Differentiation Rating Form_Standards (2a) Portfolios (accessible online) (2b) Content Differentiation Rating Form_Portfolios
5	Scoring criteria (consensus)	(1) MAP-A Test Documents a. Instructor Administration Manual b. Scorer Training Manual (pages 24-34) c. Portfolios (accessible online)

		(2) Alternate Achievement Standards (3) Rating Scale Code Descriptions – Scoring Inferences
6	Whole Test Considerations (consensus)	(1) Whole Test_Rating Forms (1 per grade) (2) Rating Scale Code Descriptions – Accessibility Dimensions (as reference) (3) NCEO Considerations Guide (as reference)

1 Rate DOK of Standards

- A. Using the 'Select Missouri Show-Me Standards' printouts, assign a depth-of-knowledge rating to the Show-Me Standards most relevant to the MAP-A. You will provide a single rating per row on the combination of knowledge standards and goals listed. First, you will rate the standards independently. Then, we will come to consensus on the ratings (3/4 majority). Use the DOK rating scale as a guide to choosing ratings.
- B. Using the A-GLEs for your content area, assign a DOK rating to each blank (non-shaded) cell on the A-AGLE DOK Rating Form. Again, we will do individual ratings followed by consensus analysis.

Decision Rule: When rating DOK of content standards, go with the **higher** level if wavering between levels.

2 Rate the Alignment Dimensions of the A-GLEs (Rate ONLY Grades 3 through 12)

Using the rating scale codes on page 2 of the Rating Scale Code Descriptions, evaluate each individual API on all of the dimensions (columns) in the form. Repeat these same tasks for each grade (5, 8, and 11). All ratings will occur independently from other reviewers (if you have a question, ask a facilitator).

- A. Academic: Determine whether the content listed in the API is academic in nature (as opposed to functional or foundational skills).
- B. Standard Match: Referring to the standards listed at the bottom of the A-GLE handout, determine whether each API matches the standards listed by indicating 'Y' (yes) or 'N' (no). If the API does not match, please enter 'N' in this column. Then, refer to the full Show-Me Standards (copies available) to determine if another standard is appropriate and enter the standard number in Notes/Comments.
- C. Content Centrality: Indicate *how well* you think that the API actually links to listed standard on content. Please use a code of '1' (No Link) only when the API does not link to any standard at all.
- D. Performance Centrality: Determine the extent to which the API measures student performance expected in the standard. NOTE: If you chose a different standard, evaluate the API against that standard instead of the one listed.
- E. Age Appropriate: Evaluate whether the API is appropriate for the age (grade) at which the content is measured.
- F. Level of Communication: Evaluate the level of communication required to demonstrate content knowledge. NOTE: Please consider the lowest functioning student who could access this API.
- G. Accessibility: Evaluate the degree of accessibility of this API for various disability groups. If the statement is accessible to all groups, enter a 'Y' (yes). If you think that the content is NOT accessible to some groups, enter 'N' (no) and provide an annotation in the Notes/Comments column to indicate those groups negatively affected. Use the Instructor's Administration Manual as references for your ratings.

- H. Accommodations and Supports: Determine whether the API is amenable to various forms of accommodations for students with various disabilities as well as supports needed by these students. Use the Instructor's Administration Manual and the NCEO Considerations Guide as references for your ratings.

I.

3 Rate the Alignment Dimensions of the portfolio entries

Access the portfolios online following instructions given by ARC. Your facilitator will instruct you on which sets of portfolios to begin rating. You will be rating up to 15 portfolios per grade (total of 45 portfolios across grades).

For each portfolio, examine the Task/Activity developed by the teacher relative to the API listed on the Entry - Data Form. Each portfolio contains multiple entries (4 each for Math and Comm Arts; 2 entries for Science). Rate each entry using the same rating scales as for the A-GLEs. Instructions for several scales differ; these different instructions are listed below:

- A. Academic: Determine whether the Task/Activity is primarily academic in nature.
- B. API Match: Determine whether the content/performance of the Task/Activity matches the API listed (Y' or N'). If the content does not match, enter N' in this column, and review the A-GLEs to determine if another API is appropriate; enter that API in Notes/Comments.
- C. Content Centrality: Indicate *how well* you think the Task/Activity links to API on content. Use a code of 1' (No Link) only when the Task/Activity does not link to any API at all.
- D. Performance Centrality: Determine extent that Task measures student performance expected in the API. NOTE: If you chose a different API, evaluate the Task against that choice.
- E. Age Appropriate: Evaluate whether the Task/Activity is appropriate for the age (grade) at which the content is measured.
- F. Level of Communication: Evaluate the level of communication required *for this student* to demonstrate content knowledge required by the task.
- ~~G. Accessibility: Not utilized for the portfolios.~~
- H. Accommodations and Supports: Determine whether the API is amenable to various forms of accommodations for students with various disabilities as well as supports needed by these students. Use the Instructor's Administration Manual and the NCEO Considerations Guide as references for your ratings.

4 Rate content differentiation across grades (A-GLEs and portfolios)

This task involves two separate global evaluations per grade: (1) A-GLEs, (2) portfolios. Using the Content Differentiation Rating Forms, compare the content expectations of the A-GLEs between grade levels/spans (should see increases in breadth and depth). Provide a holistic judgment about the differences found between grades using the Rating Scale Code Descriptions. Perform the same ratings on the portfolios. You should have 6 completed rating forms (3 for A-GLEs, 3 for portfolios) when you are finished. Although you will be providing global ratings, please cite evidence from specific portfolios as often as possible.

5 Rate Whole Test Considerations based on the portfolios you reviewed within a given grade. to demonstrating student knowledge

This is a global evaluation of the MAP-A. Instead of providing ratings for individual portfolios, consider your impressions of the set of portfolios you reviewed in each grade. Use a separate Whole Test Rating form for each grade. Consider each

student group who may be taking the assessment. These evaluations only require a Y (yes) or N (no) response in each of the blank cells. You may perform this task as a group (if everyone else has completed their individual ratings). You should have 3 completed ratings forms (1 per grade). Although you will be providing global ratings, please cite evidence from specific portfolios as often as possible.

6 Rate scoring criteria (evaluate scoring rubric, rules, achievement descriptors, and instructions for administration)

This is a global evaluation of the MAP-A. In addition to the portfolios, review the test documentation to get a sense of the extent to which students are capable of demonstrating independent learning AND whether it is possible to determine whether student work is independent or assisted. These documents *should* provide information about student performance, rather than assessment system or teacher performance. Refer to the Rating Scale Code Descriptions for explanation of codes. You should have 3 completed ratings forms (1 per grade). Although you will be providing global ratings, please cite evidence from specific portfolios as often as possible.

Rating Scale Code Descriptions

Depth of Knowledge (DOK) Scale (for use with the Full Content Standards, Alternate Standards, and Assessment)

Level	DOK Description
0	None (no content clearly measured; too vague)
1	Attention (touch, look, vocalize, respond, attend).
2	Memorize/recall (list, describe (facts), identify, state, define, label, recognize, record, match, recall, relate).
3	Performance (perform, demonstrate, follow, count, locate, read).
4	Comprehension (explain, conclude, group/categorize, restate, review, translate, describe (concepts), paraphrase, infer, summarize, illustrate).
5	Application (compute, organize, collect, apply, classify, construct, solve, use, order, develop, generate, interact with text, implement).
6	Analysis, Synthesis, Evaluation (pattern, analyze, compare, contrast, compose, predict, extend, plan, judge, evaluate, interpret, cause/effect, investigate, examine, distinguish, differentiate, generate).

**Content and Performance Dimensions
(for use with Alternate Content Standards and Assessment)**

Category	Code	Description
Academic	A	Academic – content includes major domains/strands in State standards.
	F	Functional (or Foundational) – primary content focuses on practical skills, such as daily living skills (e.g., hand washing) or pre-academic (e.g., orienting a book, lines/marks on a page with pencil).
	N	Neither functional nor academic.
Standard Match		See full content standards and alternate content standards.
Content Centrality	1	No link
	2	Weak link
	3	Moderate link
	4	Close link
Performance Centrality	A	All - performance expectation is identical to content standard.
	S	Some - performance expectation partially matches content standard (content standard may include two different performance expectations, such as <i>'Identify and explain'</i>).
	N	None - performance expectation is different from content standard
Age Appropriate	A	Adapted from grade-level content
	I	Inappropriate; off-grade content (should be taught/assessed at a higher or lower grade level).
	N	Neutral; content is not age-bound, but could be taught at any grade (NOT COMMON)

**Accessibility Dimensions
(for use with Alternate Content Standards and Assessment)**

Category	Code	Description
Levels of Communication	P	Pre-symbolic - student may demonstrate intentionality by showing interest, focus, or desire for a result through behavior; can use idiosyncratic gestures, sounds, or purposeful movements but no discrimination between pictures or other symbols.
	E	Early Symbolic - student demonstrates emerging knowledge of symbols with some recognition of symbol-object relationships.
	S	Symbolic - student has broad knowledge of and can communicate consistently with symbols (e.g., pictures) or words (e.g., speech, assistive technology, signs).
Access	Y	Yes, the standard is accessible to all students.
	N	No, some students cannot access the content of this standard or item (PLEASE provide annotation in Notes to explain).

Accommodation/ Supports	Y	Yes, students can access content with appropriate accommodations (e.g., audio) or supports (e.g., objects; assistive technology).
	N	No, the content is not amenable to accommodations or supports.
* Portfolio includes additional code		

**Content Differentiation Across Grades
(for use to compare between grades for Alternate Standards and for
Assessment)**

- (a) **broader**—higher-grade standards or items reflect broader application of target skill/knowledge;
- (b) **deeper**—higher-grade standards or items reflect deeper mastery of the target skill/knowledge
- (c) **prerequisite**—lower-grade standards or items reflects a different by prerequisite skill for mastery of the higher grade standard;
- (d) **new**—the higher-grade has a new skill or knowledge unrelated to skill/knowledge covered at prior grades; and
- (e) **identical**—higher-grade standards or items appear identical to one of the lower-grade standards.

Scoring Inferences

(Use to evaluate level of independence evident in student work provided in portfolios)

Degree of Inference about Student Learning (based on scoring for each AA item or found in the standards setting information)

Criterion	High Student Inference Can clearly infer student showed learning	Low Student Inference Student performance mixed with educator performance	No Student Inference Can clearly infer student did not have to show any learning/ Teacher or program performance rated ("Raggedy Andy" would pass)	Rationale for Rating (provide where evidence found)
Level of accuracy	High level of accuracy (If one response; response is correct. If multiple responses, above 90% correct)	Lower level of accuracy or accuracy intermixed with teacher assistance to extent difficult to determine what student did.	Does not have to get items correct to receive credit.	
Level of independence	Only independent response receives credit (Students may receive a verbal question/ direction to respond but not told what response to make)	Credit given for responses in which student performs either without guidance after told or shown the exact response to make (verbal, model prompts, scaffolding) or are done after shown/ told exact response to make and also given some guidance to make the response (partial physical)	Credit given for responses made with hand over hand assistance	
New learning (important to AA because alternate achievement is not as clear as grade level)	Baseline or pretest provides support that this is new learning OR One time performance but clear differentiation of AA items by grade level (criteria 5)	One time performance AND grade level differentiation of AA items was not clear (criteria 5)	No baseline, pretest, and weak differentiation across grade level AA items suggest student could achieve proficiency by making same response year after year (criteria 5).	

Criterion	High Student Inference Can clearly infer student showed learning	Low Student Inference Student performance mixed with educator performance	No Student Inference Can clearly infer student did not have to show any learning/ Teacher or program performance rated ("Raggedy Andy" would pass)	Rationale for Rating (provide where evidence found)
Generalization across people and settings (Note: this is less important than conceptual generalization)	Tasks are demonstrated across people or settings for full credit	At least some tasks are demonstrated across more than one person or setting	Task is only demonstrated with one person in one setting	
Generalization across materials and activities (conceptual generalization)	Tasks are demonstrated across materials and activities or all standards have more than one task	At least some tasks are demonstrated across materials or activities; or there is more than one task for some standards	Task is only demonstrated with one specific material and activity; there is only one task per standard	
Standard Setting	Standard set for proficiency is based on independent student performance and high level of accuracy	Standard set for proficiency will require student show some independent responding and respond correctly above chance level	Standard set for proficiency is so low students could meet it with either chance responding or prompting that gives student the answer	
Program Quality Indicators	If program quality indicators are used, they are not factored into student score	If program quality indicators are used, they have minimal impact on student score (e.g., small portion of rubric)	Student score is heavily influenced by program quality indicators in rubric	