

## **Independent Alignment Review of the Science Minnesota Test of Academic Skills (MTAS)**

Leslie R. Taylor  
Arthur A. Thacker  
Hilary L. Campbell  
Emily R. Dickinson  
Lisa E. Koger  
Richard C. Deatz  
R. Gene Hoffman

*Prepared for:* Minnesota Department of Education  
1500 Highway 36 West  
Roseville, MN 55113

*Prepared under:* Contract No:

September 26, 2008



## **Independent Alignment Review of the Science Minnesota Test of Academic Skills (MTAS)**

Leslie R. Taylor  
Arthur A. Thacker  
Hilary L. Campbell  
Emily R. Dickinson  
Lisa E. Koger  
Richard C. Deatz  
R. Gene Hoffman

*Prepared for:* Minnesota Department of Education  
1500 Highway 36 West  
Roseville, MN 55113

*Prepared under:* Contract No:

September 26, 2008



## EXECUTIVE SUMMARY

### *Scope of Work*

The Minnesota Department of Education (MDE) requested an external independent alignment study (review and analysis) of the Science Minnesota Test of Academic Skills (MTAS) for students with significant cognitive disabilities in grades 5, 8, and high school. The high school Science MTAS should be administered when the student receives Life Science instruction; it is the role of the IEP team to determine the most appropriate year for a student to complete the Science MTAS. Specifically, MDE wanted an evaluation of the alignment between the Science MTAS grade-level assessment, the alternate or extended content standards (referred to in Minnesota as Essence Statements<sup>1</sup>), the Minnesota Academic Standards<sup>2</sup>, and newly constructed alternate achievement standards. Minnesota uses the Science MTAS test in the federal and state accountability programs. The Human Resources Research Organization (HumRRO) was awarded a contract to conduct this alignment study, and work began on June 2, 2008.

MDE 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 (NCLB) Act of 2001. Alternate assessments are included in this requirement. 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 its AYP calculations by setting different performance expectations for up to 1% of the student population. As a result, states can develop alternate or extended content standards, achievement standards, and assessments designed to demonstrate more fairly 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 does at least link to the grade level expectations, although the breadth and depth of these expectations can be reduced (USDE, 2005).

---

<sup>1</sup> Essence Statements and Alternate Achievement Standards can be found in the MTAS Test Specifications: [http://education.state.mn.us/MDE/Accountability\\_Programs/Assessment\\_and\\_Testing/Assessments/MTAS/MTAS\\_Test\\_Specifications/index.html](http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Test_Specifications/index.html).

<sup>2</sup> Minnesota Academic Standards can be found at [http://education.state.mn.us/MDE/Academic\\_Excellence/Academic\\_Standards/index.html](http://education.state.mn.us/MDE/Academic_Excellence/Academic_Standards/index.html).

## ***Methodology***

Three different types of alignment evaluations were performed using a 2008 Science MTAS test form: (a) the extended standards (Essence Statements) to the full content standards (Minnesota Academic Standards), (b) the Science MTAS assessments to the Essence Statements, and (c) the assessments to the alternate achievement standards. Alignment evaluations (a) and (b) involved a review of the performance tasks by current and recently retired Minnesota educators highly familiar with the content standards and the assessment. For the third evaluation (c), HumRRO compared the difficulty of the performance tasks to the established achievement (proficiency) standards and cut scores. This last review did not involve external panelists.

### **Review of Content Alignment and Accessibility**

For the evaluations of the Essence Statements and assessments, HumRRO convened three separate panels (five panelists each) to review the grades 5, 8, and high school science assessments<sup>3</sup>. Each panelist carried out two primary tasks by performing multiple ratings: (a) comparison of the grade span Essence Statements to the Minnesota Academic Standards for science, and (b) comparison of the Science MTAS performance tasks (per grade test) to the Essence Statements. The purpose of these tasks was to evaluate the content alignment of the Essence Statements and assessments relative to the full Minnesota Academic Standards. In addition, panelists reviewed the content and performance accessibility of the Essence Statements and assessments to the population of students for whom the alternate assessment was designed.

HumRRO developed the review panels with the assistance of MDE and Pearson Educational Measurement. Panelists were recruited by Pearson from their database of Minnesota educators. Every effort was made to produce panels consisting of teachers reflecting the population of students who take the assessments. Panels were convened in facilities procured through MDE. HumRRO directed the actual reviews independently of MDE and Pearson.

HumRRO used the Links for Academic Learning alignment method (referred to as the Links method in this report) developed by the National Alternate Assessment Center 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):

---

<sup>3</sup> Sample science performance tasks can be found through the MDE website: [http://education.state.mn.us/MDE/Accountability\\_Programs/Assessment\\_and\\_Testing/Assessments/MTAS/MTAS\\_Item\\_Samplers/index.html](http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Item_Samplers/index.html).

**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, math, science).

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

**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.

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

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

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

**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 Criterion 4 above, we refer to the "Webb alignment indicators". Dr. Norman Webb developed an alignment procedure involving an evaluation of the assessment to the content standards using four statistics (2005). These statistics indicate how well an assessment covers the content standards in terms of content breadth and depth (2005). 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 even for an alternate assessment. Thus, the Links 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., standard, benchmark) 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 consistency – compares the cognitive complexity ratings of the items with the complexity ratings of each content standard.

The outcomes of the analyses on the Links criteria and Webb alignment indicators are evaluated against decision rules to judge their acceptability.

## **Review of Performance Alignment**

For the review of performance alignment, HumRRO analyzed the science tasks for each grade's assessment relative to the achievement standards used to make student classifications. Students are classified into one of four levels of performance established by Minnesota based on their test scores: (a) Exceeds the Standards, (b) Meets the Standards, (c) Partially Meets the Standards, or (d) Does Not Meet the Standards. Because the outcome of student performance will be included in NCLB accountability decisions, it is important to confirm that the assessments are functioning as intended, which is discriminating among students within the range of the established assessment cut scores.

### ***Summary Alignment Results***

#### **Key Findings and Conclusions**

The results of the alignment reviews provide positive support overall for the content validity of the Science MTAS assessment based on several outcomes. First, panelists found the majority of grade-span Essence Statements to be linked adequately to the full Minnesota Academic Standards in content breadth and depth. Second, all MTAS performance tasks were rated as matched to content expectations, and each assessment measured a reasonable range of content at varying levels of cognitive complexity. Finally, the performance alignment review of achievement standards indicated that assessments can discriminate among students in the range of the established achievement levels.

#### **Essence Statements to Minnesota Academic Standards for Science**

Table 1 displays the overall conclusions regarding content alignment between the Essence Statements and Minnesota Academic Standards for science. These judgments are based on whether the Essence Statements achieved acceptable levels of linkage with the full content standards for each grade test. The minimum level for each of the criteria in Table 1 is 90%.

- High linkage - most 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 Essence Statements to Science Minnesota Academic Standards on Links Criteria 2, 3, and 5**

Grade Level Tests	Criterion 2	Criterion 3		Criterion 5
	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	Partial	Partial	Partial
8	High	Partial	Partial	Partial
High School	High	High	High	Partial

As with most alignment reviews, some areas of weakness were identified. These outcomes warrant further explanation. Specifically, while all of the Essence Statements were rated as linked to a Minnesota benchmark, some Essence Statements for grades 5 and 8 did not sufficiently link to the central grade level content.

Concerning content differentiation, most panelists agreed that the expectations of the Essence Statements require students to demonstrate more advanced knowledge as grade levels increase. However, the high school grade-span Essence Statements showed inconsistent increases in content expectations relative to earlier grades. In particular, panelists determined that these Essence Statements as a whole did not cover a broader range of content compared to lower grade expectations, and they considered some of the content to be too similar. Panelists found that the high school Essence Statements showed evidence of more complex knowledge expectations (deeper) and *some* new information not presented at earlier grades.

Table 2 displays the overall conclusions on content accessibility pertaining to Performance Accuracy (content accessibility) for the Essence Statements. For this criterion, conclusions reflect overall judgments of acceptability<sup>4</sup>.

- 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%)

<sup>4</sup> Adapted from universal design ratings used by the National Center on Educational Outcomes (NCEO). See Thompson et al. (2005).

**Table 2. Summary Conclusions on Performance Accuracy (Links Criterion 7) of Essence Statements for Science**

Criterion 7		
Grade Level Tests	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	Good	Acceptable
8	Acceptable	Excellent
High School	Excellent	Excellent

Panelists for the high school science assessment considered all of the Essence Statements to be appropriate for students at different communication and ability levels. The results for the grades 3-5 Essence Statements indicated mixed ratings on performance accuracy. For example, panelists mostly agreed that students of all levels of communication can access the content and demonstrate knowledge for these Essence Statements. However, they felt that some types of students with particular disabilities may be disadvantaged by some of the content expectations. Specifically, panelists indicated in their comments that visually impaired students may have difficulty effectively demonstrating knowledge, and that there even could be safety concerns for these students in demonstrating the following two science Essence Statements:

**Strand I – History and Nature of Science**

**MCA-II Benchmark: 3.I.B.2 and 5.I.B.1**

**MTAS Essence Statement**

The student will identify tools appropriate for a given scientific investigation.

**Strand II – Physical Science**

**MCA-II Benchmark: 4.II.A.1**

**MTAS Essence Statement**

The student will identify changes in states of matter (solid, liquid, and gas).

**Science MTAS Tasks to Essence Statements**

Table 3 provides summary conclusions on the alignment of the MTAS assessments to Essence Statements.

- High linkage - most 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 MTAS Assessments to Essence Statements for Links Criteria 1, 2, 3, 4, and 5**

	Criterion 1	Criterion 2	Criterion 3		Criterion 4		Criterion 5
Grade Level Tests	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 Essence Statements?	Does the performance of task link to expectations of the Essence Statements?	Do the tasks assess students at the appropriate breadth of knowledge? <sup>a</sup>	Do the tasks assess students at the appropriate depth of knowledge? <sup>b</sup>	Do the assessments show appropriate increases between grade levels?
5	Partial	High	Partial	High	High	Partial	Partial
8	High	High	High	High	High	Partial	Partial
High School	High	High	Partial	High	High	High	High

<sup>a</sup> This conclusion is based on a summary judgment across the Webb statistics of Categorical Concurrence, Range of Knowledge, and Balance of Knowledge. It is still important to consider each of the criteria separately as well.

<sup>b</sup> This conclusion is based on the results from the DOK consistency analyses.

Table 4 includes results relative to Criteria 6 and 7 of the Links 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 alignment<sup>5</sup>.

- 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 Performance Accuracy (Links Criterion 7) of Science MTAS Assessments**

	Criterion 6	Criterion 7		
Grade Level Tests	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	Acceptable	Questionable	Acceptable
8	Questionable	Questionable	Excellent	Excellent
High School	Acceptable	Excellent	Excellent	Excellent

### Performance Alignment

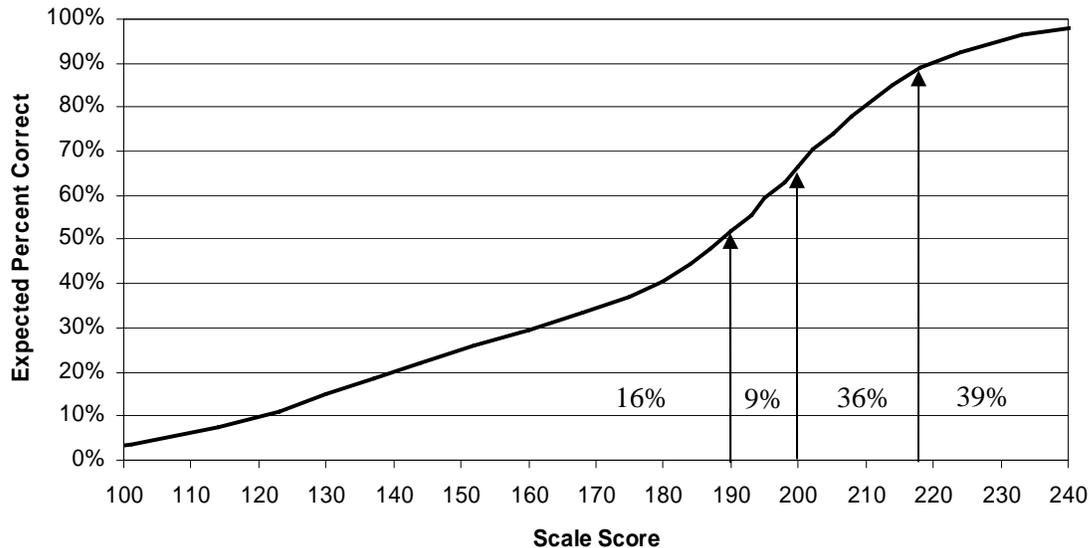
Figure 1 demonstrates the results from the review of the grade 5 Science MTAS with respect to its achievement standards. Test functioning is depicted by a “test characteristic curve” that describes the Item Response Theory-based relationship between achievement and test performance. The two cut scores between the three lowest categories are within the proportion of the curve that shows the strongest relationship between achievement and percentage of items correct. The third cut score between “meets” and “exceeds,” however, is near the high end of the TCC where discrimination among students is less strong.

In terms of Minnesota’s AYP scoring, discrimination between these top two categories is less important than discrimination among the other three categories. Increasing the overall difficulty of the assessment by replacing some of the easier tasks with more difficult tasks may improve the “meets” versus “exceeds” discrimination; however, if not done carefully, it could reduce the assessment’s ability to discriminate among the three lower categories that are used to determine schools’ AYP scores. The

<sup>5</sup> 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.

MTAS assessments must cover a large range of student abilities, which creates an equally large challenge. Therefore, this report will stop short of recommending a change in task difficulty based on these results. There are simply too many other considerations. Our conclusions for grade 8 and high school Science MTAS are the same.

### Science MTAS - Grade 5



**Figure 1. Alignment of achievement levels and Grade 5 Science MTAS test functioning.**

## Recommendations

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

### Essence Statements for Science

- (1) **Review the content differentiation between grade-span Essence Statements.** Each set of grade-span content expectations demonstrated some need for greater differentiation in content at increasingly higher grade levels. While panelists found broadening and deepening of knowledge expectations, increases were very limited in some cases. Minnesota may find this recommendation daunting because it would seem to require re-writing (and, hence, re-approval by the State) the Essence Statements. However, it may be that the current Essence Statements can be better differentiated by adding to the content limitations and/or including examples of how students might demonstrate knowledge differently at higher grade levels. Some states include examples of performance activities for each content expectation per grade level.

- (2) **Review the access points for each of the grade-span Essence Statements.** None of the Essence Statements were rated as highly exclusive – most were at least acceptable if not quite good in allowing student access and demonstration of knowledge. However, since student access is such a critical issue, we suggest that Minnesota re-examine the Essence Statements for grades 5 and 8 in particular. Such a task may involve additional bias reviews, or, as noted above, greater explication (content limitations, examples) within the Test Specifications document of how teachers and test administrators might make these content expectations more appropriate.

### **MTAS Performance Tasks**

It is important to note that no panelists considered any of the tasks to display serious flaws that would warrant complete replacement of tasks. Instead, ratings and comments by panelists point to issues that could be improved upon for better student access.

- (1) **Review some performance tasks for the grade 5 and high school assessments for clarity in targeted content (content centrality).** While panelists agreed with the test contractor on the target of assessment in most cases, panelists also indicated that some (approximately 3 to 4) performance tasks did not always measure this content well. Ratings on these tasks also correspond with lower ratings on capacity for demonstrating achievement accessibility in many cases. These combined outcomes suggest that certain performance tasks may require additional review by content and special education experts. Based on panelists' comments, such a review may only require edits to task presentation or response card options, as opposed to a complete revision of the topic of the performance task or target of assessment.
- (2) **Improve the ability of the Science MTAS assessments to accurately demonstrate student knowledge.** This recommendation reflects the combined outcomes from student inference on achievement and performance accuracy based on accessibility. For the grades 5 and 8 assessments in particular, panelists found that some tasks do not allow for clear inference about student learning, which they partly attributed to limitations in access to certain student groups. One issue in particular noted by panelists is the fact that the Science MTAS assessment does not include a pretest, or baseline, which raised some concerns due to the fact that the science assessments cover multiple grade content.

A second concern among panelists focused on the option for test administrators to apply a wide range of alternate materials for modification, thus potentially reducing standardization. We agree with the latter point in part, but with the recognition that alternate assessments should allow for reasonable modification. Such comments by panelists seem to reflect concern that test administrators or teachers may not be well versed, or comfortable, with what counts as appropriate modifications. More direction as part of training on test administration may be appropriate.

---

**INDEPENDENT ALIGNMENT REVIEW OF THE SCIENCE MINNESOTA TEST OF  
ACADEMIC SKILLS (MTAS)**

**TABLE OF CONTENTS**

<b>Chapter 1 Introduction .....</b>	<b>1</b>
<b>Chapter 2 Alignment Study Design and Methodology .....</b>	<b>3</b>
Alignment of Assessments and Standards on Content and Performance.....	3
<i>Content Alignment and Accessibility</i> .....	3
<i>Performance Alignment</i> .....	6
Scope of Alignment Evaluations for Science MTAS Assessments .....	7
<i>Review of Content Alignment and Accessibility</i> .....	7
<i>Review of Performance Alignment</i> .....	11
<b>Chapter 3 Results: Essence Statements and Minnesota Academic Standards ..</b>	<b>13</b>
Results on Essence Statements based on Links Criteria.....	13
Inter-Rater Reliability .....	22
Summary and Discussion of Essence Statements and Minnesota Academic Standards .....	23
<b>Chapter 4 Results: Science MTAS Tasks and Essence Statements.....</b>	<b>27</b>
Results on Science MTAS Tasks based on Links Criteria .....	27
Reliability Results.....	39
<i>Inter-Rater Reliability</i> .....	39
<i>Panelist-Test Developer Analyses</i> .....	40
Summary and Discussion of Science MTAS Tasks and Essence Statements .....	41
<b>Chapter 5 Results: Science MTAS Tasks and Alternate Achievement Standards</b>	<b>45</b>
Summary and Discussion of Science MTAS Tasks and Alternate Achievement Standards .....	48
<b>Chapter 6 Summary and Recommendations .....</b>	<b>49</b>
<b>References .....</b>	<b>52</b>
<b>Appendix A Webb Alignment Results per Grade Level Assessment .....</b>	<b>A-1</b>
<b>Appendix B Summary of Panelist Comments on Essence Statements and Performance Tasks .....</b>	<b>B-1</b>
<b>Appendix C Sample Alignment Review Materials .....</b>	<b>C-1</b>

## TABLE OF CONTENTS (CONTINUED)

### List of Tables

Table 2.1 Professional and Demographic Characteristics of Science MTAS Alignment Panelists.....	8
Table 2.2 Characteristics of the Science MTAS Tests Reviewed.....	9
Table 3.1 Mean Number of Essence Statements Rated as Age Appropriate.....	14
Table 3.2 Mean Number of Essence Statements at Various Levels of Content Centrality .....	15
Table 3.3 Percentages of Essence Statements at Same, Lower, Higher Levels of Complexity Compared to Related Benchmarks.....	16
Table 3.4 Mean Number of Essence Statements at Various Levels of Performance Centrality .....	17
Table 3.5 Consensus Ratings on Content Differentiation between Grade Spans of Science MTAS Essence Statements.....	18
Table 3.6 Mean Number of Essence Statements Rated at Each Level of Symbolic Communication.....	20
Table 3.7 Mean Number of Essence Statements Rated as Accessible to All Students.....	20
Table 3.8 Inter-Rater Agreement on Ratings for Essence Statements per Grade Level .....	23
Table 3.9 Summary Conclusions on Alignment of Essence Statements to Minnesota Academic Standards for Science on Links Criteria 2, 3, and 5.....	24
Table 3.10 Summary Conclusions on Performance Accuracy (Links Criterion 7) of Essence Statements for Science .....	25
Table 4.1 Mean Number of Tasks Rated as Academic by Panelists .....	27
Table 4.2 Mean Percentage of Tasks at Various Levels of Age Appropriateness .....	28
Table 4.3 Mean Number of Tasks Linked to Essence Statements.....	29
Table 4.4 Mean Percent of Items at Various Levels of Content Centrality .....	29
Table 4.5 Mean Percent of Tasks at Various Levels of Performance Centrality.....	30
Table 4.6 Summary of Categorical Concurrence Results for Science MTAS by Grade Level .....	31
Table 4.7 Mean Percentage of Items at Various Levels of DOK .....	32
Table 4.8 Summary of Depth-of-Knowledge Results for Science MTAS by Grade Level .....	33

---

**TABLE OF CONTENTS (CONTINUED)**

Table 4.9 Summary of Range-of-Knowledge Results for Science MTAS by Grade Level .....	34
Table 4.10 Summary of Balance-of-Knowledge Representation Results for Science MTAS by Grade Level .....	35
Table 4.11 Consensus Ratings on Content Differentiation between Grade Level Science MTAS Assessments.....	35
Table 4.12 Degree of Inference Evident on Student Learning in Science MTAS Assessments.....	37
Table 4.13 Mean Percentage of Tasks at Various Levels of Symbolic Communication .....	38
Table 4.14 Mean Numbers of Tasks Rated Accessible to Students.....	38
Table 4.15 Mean Number of Tasks Amenable to Modifications or Supports.....	39
Table 4.16 Inter-Rater Agreement on Panelists' Ratings of Science MTAS Tasks per Grade Level.....	40
Table 4.17 Percentage Agreement between Panelists and Pearson on Assessment Target for Science MTAS Tasks.....	40
Table 4.18 Summary Conclusions on Alignment of Science MTAS Assessments to Essence Statements for Links Criteria 1, 2, 3, 4, and 5 .....	42
Table 4.19 Summary Conclusions on Accessibility (Links Criteria 6 and 7) of Science MTAS Assessments .....	43
Table A-1. Categorical Concurrence for Science MTAS, Grade 5: Mean Number of Performance Tasks per Strand.....	A-1
Table A-2. Categorical Concurrence for Science MTAS, Grade 8: Mean Number of Performance Tasks per Strand.....	A-1
Table A-3. Categorical Concurrence for Science MTAS, High School: Mean Number of Performance Tasks per Strand.....	A-2
Table A-4. Depth-of-Knowledge Consistency for Science MTAS, Grade 5: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives .....	A-2
Table A-5. Depth-of-Knowledge Consistency for Science MTAS, Grade 8: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives .....	A-3
Table A-6. Depth-of-Knowledge Consistency for Science MTAS, High School: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives .....	A-3
Table A-7. Range-of-Knowledge for Science MTAS, Grade 5: Mean Percent of Essence Statements per Strand Linked with Performance Tasks.....	A-4

## TABLE OF CONTENTS (CONTINUED)

Table A-8. Range-of-Knowledge for Science MTAS, Grade 8: Mean Percent of Essence Statements per Strand Linked with Performance Tasks.....	A-4
Table A-9. Range-of-Knowledge for Science MTAS, High School: Mean Percent of Essence Statements per Strand Linked with Performance Tasks .....	A-5
Table A-10. Balance-of-Knowledge Representation for Science MTAS, Grade 5: Mean Balance Index per Strand.....	A-5
Table A-11. Balance-of-Knowledge Representation for Science MTAS, Grade 8: Mean Balance Index per Strand.....	A-6
Table A-12. Balance-of-Knowledge Representation for Science MTAS, High School: Mean Balance Index per Strand.....	A-6
Table B-1. Summary of Repeated Panelist Comments on Science MTAS Essence Statements .....	B-1
Table B-2. Summary of Repeated Panelist Comments on Science MTAS Performance Tasks .....	B-1

## List of Figures

Figure 5-1. Alignment of achievement levels for Grade 5 Science MTAS.....	46
Figure 5-2. Alignment of achievement levels for Grade 8 Science MTAS.....	47
Figure 5-3. Alignment of achievement levels for High School Science MTAS.....	47

## INDEPENDENT ALIGNMENT REVIEW OF THE SCIENCE MINNESOTA TEST OF ACADEMIC SKILLS (MTAS)

### Chapter 1 Introduction

The Minnesota Department of Education (MDE) requested an external independent alignment study (evaluation/analysis) of the Science Minnesota Test of Academic Skills (Science MTAS) for students with significant cognitive disabilities in grades 5, 8, and high school. The high school Science MTAS should be administered when the student receives Life Science instruction; it is the role of the IEP team to determine the most appropriate year for a student to complete the Science MTAS. Specifically, MDE wanted an evaluation of the alignment between the Science MTAS grade-level assessment, the extended content standards (or Essence Statements<sup>6</sup>), the Minnesota Academic Standards<sup>7</sup>, and newly constructed alternate achievement standards. Minnesota uses the Science MTAS test in the federal and state accountability programs. The Human Resources Research Organization (HumRRO) was awarded a contract to conduct this alignment study, and work began on June 2, 2008.

MDE requested the alignment study to meet both state and federal requirements. The federal requirement of the U.S. Department of Education (USED) stems from the No Child Left Behind (NCLB) Act of 2001. NCLB challenges each state to establish a coherent assessment system based on solid academic standards. This law calls for states to provide independent evidence of the validity of 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 this requirement. 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”

---

<sup>6</sup> Essence Statements and Alternate Achievement Standards can be found in the MTAS Test Specifications: [http://education.state.mn.us/MDE/Accountability\\_Programs/Assessment\\_and\\_Testing/Assessments/MTAS/MTAS\\_Test\\_Specifications/index.html](http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Test_Specifications/index.html).

<sup>7</sup> Minnesota Academic Standards can be found at [http://education.state.mn.us/MDE/Academic\\_Excellence/Academic\\_Standards/index.html](http://education.state.mn.us/MDE/Academic_Excellence/Academic_Standards/index.html).

(U.S. Department of Education, 2005). This rule 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 demonstrate more fairly 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 does at least link to the grade level expectations, although the breadth and depth of these expectations can be reduced (USDE, 2005).

### ***Organization and Contents of the Report***

This report contains six chapters. Chapter 2 explains alignment methodologies, including general methods used to evaluate alignment of alternate assessments. Subsequent chapters provide alignment results for comparisons between the components of the assessment system: (a) Chapter 3 presents results of the alignment comparison between the Essence Statements and the Minnesota Academic Standards; (b) Chapter 4 presents results of the content review of the Science MTAS tasks relative to the Essence Statements; (c) Chapter 5 includes an analysis of the Science MTAS tasks against the newly developed alternate achievement standards and cut scores; and (d) Chapter 6 provides recommendations for MDE to strengthen alignment of the Science MTAS assessment over time.

Additional information is provided in the appendices to this report. Appendix A contains tables providing more detail on the content alignment results for the grade-level test forms. Appendix B includes a summary of panelists' comments on their ratings based on the type of comment provided. Appendix C provides examples of rating forms and training materials used in the alignment workshops.

## Chapter 2 Alignment Study Design and Methodology

In this section, we discuss key concepts related to alignment research, followed by a description of the alignment evaluations and methods used as part of the Minnesota study.

### ***Alignment of Assessments and Standards on Content and Performance***

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

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

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

These questions more or less can be grouped into two categories – content alignment and performance alignment. However, all alignment evaluations tie back to the state content standards.

### **Content Alignment and Accessibility**

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. Ratings are made by education experts and then analyzed statistically 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 Alignment Method.***

For the current alignment study, HumRRO used the Links for Academic Learning alignment method (referred to in this report as Links) 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 seven criteria. These criteria are listed below (adapted from Flowers et al, 2007):

**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, math, science).

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

#### **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.

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

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

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

**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 Links method is appropriate for alignment of assessments to standards and alignment of extended standards to full content standards. The review of assessments to standards, such as the MTAS assessment to the Essence Statements for science, includes all of the Criteria 1 through 7. However, only Criteria 2, 3, 4, 5, and 7 can be

applied to a review of extended standards. Criterion 1 is intended to evaluate the assessment tasks as to whether they are written as academic in content, while Criterion 6 is intended to evaluate the measurement accuracy of the assessment.

### ***Webb Alignment Method.***

Under Criterion 4 of the Links method, we refer to “Webb alignment indicators”. Dr. Norman Webb 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 (2005). Webb’s method generally has been applied to 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 even for an alternate assessment. Thus, the Links method includes the Webb alignment indicators.

The Webb alignment method has been used extensively to conduct alignment reviews of regular assessments to state content standards (e.g., Webb, 1997; 1999; 2005), and his approach is supported by the Council of Chief State School Officers (CCSSO). The Webb approach includes four alignment indicators linked with statistical procedures to assess how well the assessment matches individual portions of the standards documents. The four alignment criteria are (a) categorical concurrence, (b) depth-of-knowledge consistency, (c) range-of-knowledge correspondence, (d) and balance-of-knowledge representation.

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

***Depth of Knowledge*** (DOK) measures the type of cognitive processing required by items and content standards. For example, is a student expected to simply identify or recall basic facts, or is the student expected to use reasoning by manipulating information or strategizing? In mathematics, a student may be asked to identify the appropriate use of a decimal among several answer choices. This task should be less complex than trying to explain the concept of a decimal and how and why it can be moved. In English-language arts, asking a student to identify Greek mythology requires less processing compared with asking a student to use knowledge of Greek mythology to understand the origin and meaning of new words.

The purpose of using DOK as a measure of alignment is to determine whether a test item (or performance task) and corresponding standard are both written at the same level of cognitive complexity. Reviewers make two separate judgments about cognitive complexity, one for the standard and one for the item. These two judgments are compared to determine whether the item is written at the same level as the standard to which it is linked. Webb refers to his comparison as *depth-of-knowledge consistency*.

Another measure examines the ***range-of-knowledge correspondence*** between the assessment and content standards. The range-of-knowledge measure examines in greater detail the breadth of knowledge represented by test items. Categorical concurrence simply notes whether a sufficient number of items on the test covers each general content topic (individual strands). However, states generally lay out more specific *content objectives*, or standards, under each strand. The range indicates the number of content objectives assessed by items.

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

### **Performance Alignment**

For the review of performance alignment, HumRRO analyzed the science tasks for each grade assessment with respect to the achievement standards. Students are classified into one of four levels of performance established by Minnesota based on their test scores: (a) Exceeds Standards, (b) Meets Standards, (c) Partially Meets the Standards, or (d) Does Not Meet the Standards. Because the outcome of student performance will be included in NCLB accountability decisions, it is important to confirm that the Science MTAS assessments can differentiate between the performance categories above.

The cut scores themselves were also graphically presented on Item Response Theory (IRT) Test Characteristic Curves (TCC). These curves relate the IRT-derived achievement scale to the expected percentage of items answered correctly. The TCCs should show strong upward trends in the regions of the cut scores.

### ***Scope of Alignment Evaluations for Science MTAS Assessments***

Three different types of alignment evaluations were performed for this Minnesota study: (a) the extended standards (Essence Statements) to the full content standards (Minnesota Academic Standards) for science, (b) the Science MTAS assessments to the Essence Statements, and (c) the assessments to the alternate achievement standards. Alignment evaluations (a) and (b) involved a review of the performance tasks by current and recently retired Minnesota educators highly familiar with the content standards and the assessment. For the third evaluation (c), HumRRO compared the difficulty of the performance tasks to the established achievement (proficiency) standards and cut scores. This last review did not involve external panelists.

### **Review of Content Alignment and Accessibility**

For the evaluations of the Essence Statements and assessments identified as (a) and (b) above, HumRRO convened panels of Minnesota educators. Each panelist carried out two primary alignment tasks by performing multiple ratings: (a) comparison of the grade-span Essence Statements to the Science Minnesota Academic Standards, and (b) comparison of the Science MTAS performance tasks (per grade test) to the Essence Statements. The purpose of these tasks was to evaluate the content alignment of the Essence Statements and assessments relative to the full Minnesota Academic Standards. In addition, panelists reviewed the content and performance accessibility of the Essence Statements and assessments relative to the population of students for whom the alternate assessment was designed. HumRRO applied the Links method to conduct these reviews.

#### ***Panelists.***

For the evaluations of the Essence Statements and assessments, HumRRO convened three separate panels (five panelists each) to review the grades 5, 8, and high school science assessments. HumRRO developed the review panels with the assistance of MDE and Pearson Educational Measurement. Panelists were recruited by Pearson from their database of Minnesota educators. Every effort was made to produce panels consisting of teachers reflecting the population of students who take the assessments. Panels were convened in facilities procured through MDE. HumRRO directed the actual reviews independently of MDE and Pearson. Table 2.1 presents the characteristics of the panelists per grade-level Science MTAS assessment.

**Table 2.1 Professional and Demographic Characteristics of Science MTAS Alignment Panelists**

Professional Position	Number of Panelists	Average Years of Experience <sup>a</sup>	Special Certifications	Region of Origin in Minnesota			Gender		Ethnicity				
				7County Metro	Greater Minnesota	MPLS/ St Paul	M	F	White, Non-Hispanic	Hispanic	Black, Non-Hispanic	Asian/Pacific Islander	American Indian/Alaskan Native
<b>Grade 5</b>													
Teacher	3	13.50 (n = 2)	0	2	1	0	1	2	3	0	0	0	0
Administrator	1		0	1	0	0	0	1	1	0	0	0	0
Higher Education	0		0	0	0	0	0	0	0	0	0	0	0
<b>Grade 8</b>													
Teacher	3	10.33 (n = 3)	0	1	2	0	2	1	3	0	0	0	0
Administrator	1		1	0	1	0	0	1	1	0	0	0	0
Higher Education	0		0	0	0	0	0	0	0	0	0	0	0
<b>High School</b>													
Teacher	4	9.00 (n = 3)	2	2	2	0	1	3	4	0	0	0	0
Administrator	0		0	0	0	0	0	0	0	0	0	0	0
Higher Education	0		0	0	0	0	0	0	0	0	0	0	0

<sup>a</sup> No information on experience was available for several panelists; thus, n < 4 in this column.

### ***Materials.***

Panelists evaluated the alignment of the MTAS performance tasks with the Minnesota Academic Standards and Essence Statements using forms for both the Webb and Links alignment methods. All rating forms were completed electronically in Microsoft Excel.

*Test Forms.* Reviewers evaluated the 2008 Science MTAS test forms. One form was evaluated per grade. Table 2.2 below lists the number of operational performance tasks and the number of content standards evaluated for each grade level.

***Table 2.2 Characteristics of the Science MTAS Tests Reviewed***

Grade Level	Number of Operational Items	Number of Field-Test Items	Number of Essence Statements
5	9	6	6
8	9	6	6
High School	9	6	6

*Rating Forms and Instructions.* To complete all necessary ratings for the Webb and Links alignment methods, panelists completed three rating forms individually and an additional three rating forms via group consensus (see Appendix C for samples of each). Panelists were provided instruction sheets enumerating the six 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 C). As the HumRRO staff guided them, panelists moved through each of the six tasks. Specific procedures for each task are provided in more detail below.

### ***Procedures.***

HumRRO conducted this alignment review at the Minnesota Department of Education on July 21-22, 2008. 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 introduce tasks the reviewers would be performing. Reviewers had the opportunity to practice conducting ratings during the large group session, which generated discussion about how to apply rating criteria.

Following the general introduction, panelists began working within their content groups. Science MTAS reviewers were split into three groups according to grades 5, 8, and high school tests. All groups contained four reviewers. HumRRO staff each supervised one to two groups.

Within their small groups, HumRRO staff further trained reviewers using sample standards and assessment tasks, facilitating discussion among group members, and answering questions about the alignment process. Regarding instructions on how to rate standards and items, HumRRO staff provided general suggestions and comments when appropriate; however, they emphasized to reviewers that staff would not give

explicit direction on how to rate standards or items because reviewers were valued as content experts. Each panelist received a laptop with rating forms already uploaded and formatted. HumRRO staff provided brief instructions about how to use the electronic rating forms.

After reviewing sample DOK evaluations as a group, reviewers rated the benchmarks from the Minnesota Academic Standards relevant to each grade-level test. For example, the grade 5 group reviewed the full content standards associated with grades 3, 4, and 5. 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 benchmark; a voluntary scribe within each group recorded these consensus ratings. Next, reviewers followed the same process to rate the DOK of the Essence Statements, first individually and then to consensus.

Next, reviewers rated the Essence Statements on a variety of factors, including (a) whether the benchmark listed is the best match, (b) how well the Essence Statement links to the benchmark, (c) whether the Essence Statement measures student performance of the benchmark, (d) whether the Essence Statement 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 Essence Statement is accessible to various disability groups. These ratings were conducted individually; no consensus ratings were conducted.

Reviewers then received more specific instructions for rating 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 into electronic rating forms on their laptops. The panelists rated the items on the same dimensions they rated each Essence Statement (see above). In addition, reviewers were instructed to assign a *primary benchmark* to an item based on a judgment that an item clearly measured this benchmark. Furthermore, reviewers could assign an *additional standard* only if the item seemed to assess another standard as clearly as the primary standard. 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 MTAS tests. HumRRO staff trained panelists on each task, and then the voluntary scribe from within the small group recorded the group's consensus ratings in preformatted Excel spreadsheets. The first consensus task required panelists to rate whole test barriers, or aspects of the MTAS as a whole that might prevent students with various disabilities from fully participating (with or without modifications/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.

## Review of Performance Alignment

For the review of performance alignment, HumRRO analyzed the science tasks for each grade assessment with respect to the achievement standards. Students are classified into one of four levels of performance established by Minnesota based on their test scores: (a) Exceeds the Standards, (b) Meets the Standards, (c) Partially meets the Standards, or (d) Does not meet the Standards. Because the outcome of student performance will be included in NCLB accountability decisions, it is important to confirm that the Science MTAS assessments can differentiate between the performance categories above.

The cut scores themselves were also graphically presented on Item Response Theory (IRT) Test Characteristic Curves (TCC). These curves relate the IRT-derived achievement scale to the expected percentage of items answered correctly. The TCCs should show strong upward trends in the regions of the cut scores.



## Chapter 3 Results: Essence Statements and Minnesota Academic Standards

The alternate assessment system should link to the full academic content standards on several dimensions, and it should provide appropriate access to the students for whom the alternate assessment was designed. In this chapter, we describe the results of the evaluation of the science Essence Statements compared to the Minnesota Academic Standards for science. These analyses relate to Criterion 2, 3, 5 and 7 of the Links method.

### *Results on Essence Statements based on Links Criteria*

Panelists rated the Essence Statements on a number of scales with various response options. Most results reported here refer to mean ratings on these scales. To analyze these ratings, we first counted how many Essence Statements were rated at each response option per panelist on all of the scales. From these counts, we then calculated the mean number of Essence Statements per response option (across panelists) for each rating scale. Results of these analyses are presented for each set of Essence Statements per grade span: grades 3-5, grades 6-8, and grades 9-12. Each grade span includes six Essence Statements.

Based on the Links method, most criteria include a minimum percentage of acceptable ratings on the Essence Statements to demonstrate reasonable linkage with the full content standards. Percentages are based on the mean ratings; thus, percentage totals on a rating scale per grade level may sum to greater than 100% in some cases. It is also important to keep in mind that these percentages are based on 4 panelists' ratings of 6 Essence Statements. In other words, a small number of raters evaluated a small number of Essence Statements.

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

Criterion 2 pertains to the developmental level of the content included in the Essence Statements. For this evaluation, panelists were asked whether the content of the science Essence Statements is appropriate for the age and grade level indicated. Several response options were possible:

- 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 3.1 includes the results of panelists' evaluations. Column 2 lists the rating categories, while the 'Mean' in Column 3 refers to the mean number of statements receiving that rating across panelists. Column 5 represents this same mean as a percentage of the total number of Essence statements per grade. For this criterion, at least 90% of Essence Statements should be rated as 'adapted' or 'neutral'<sup>8</sup>.

<sup>8</sup> The Links method does not specify a minimum for Criterion 2. This minimum level was established by HumRRO.

**Table 3.1 Mean Number of Essence Statements Rated as Age Appropriate**

Grade	Age-Related Content	Mean	SD	Percentage of Essence Statements per Rating
5	Adapted	2.75	1.26	46%
	Neutral	3.25	1.26	54%
	Inappropriate	0	0	0
8	Adapted	6.00	0	100%
	Neutral	0	0	0
	Inappropriate	0	0	0
High School	Adapted	6.00	0	100%
	Neutral	0	0	0
	Inappropriate	0	0	0

None of the Essence Statements was judged completely inappropriate by any raters at any grade level. For grades 6-8 and for high school, all six Essence Statements were rated as clearly adapted from appropriate grade-level content. For grades 3-5, about half of the Essence Statements were rated as age appropriate, but all panelists rated some statements as 'neutral' (Range = 2 to 5 statements).

### **Criterion 3: Standards Fidelity**

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

To meet Criterion 3, panelists were asked to provide several ratings indicating their judgments of the degree of content match between the Essence Statements and Minnesota Academic Standards for science. First, we asked panelists to provide a simple evaluation (yes or no) of whether the benchmarks listed as linked with the Essence Statements did, in fact, match. For those statements judged as matched to the designated benchmark, we then asked panelists to go further with a second rating to indicate *how well* the Essence Statement linked to the benchmark.

Concerning overall content match, panelists at each grade level rated all (100%) of the science Essence Statements as matched to the primary benchmark<sup>9</sup>. However, at least half of the panelists found additional benchmarks listed to be beyond the scope of the Essence Statement, particularly for the grades 3-5 benchmarks. Panelists' individual comments regarding these specific Essence Statements and corresponding benchmarks is found in Appendix C.

For the second evaluation, panelists reviewed each grade-span Essence Statement for the degree of link to the central content targeted by the benchmarks. In

<sup>9</sup> Some Essence Statements are linked to more than one benchmark in the MTAS Test Specifications for Science (MDE, January 2008). The first benchmark listed was considered the primary benchmark.

this case, panelists used the following 4-point scale to determine how well the Essence Statement reflects the benchmark content:

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

For Criterion 3, at least 90% of extended standards should be rated as 'moderate' or 'close' to the full standards. Table 3.2 shows that panelists found many of the Essence Statements to link sufficiently ('moderate' or 'close' link) with the benchmarks, and no content statements were rated as entirely different from the benchmarks. However, one to two panelists per grade determined that several Essence Statements deviated considerably from the central content targeted in their corresponding benchmarks. The Essence Statements for grades 5 and 8 in particular may not link sufficiently to the benchmarks because only 75% were rated as linked to the target content of the Minnesota Academic Standards.

**Table 3.2 Mean Number of Essence Statements at Various Levels of Content Centrality**

Grade	Content Centrality Rating	Mean	SD	Percentage of Essence Statements per Rating
5	No link	0	0	0
	Weak link	1.50	0.58	25%
	Moderate link	1.75	0.50	29%
	Close link	2.75	0.96	46%
8	No link	0	0	0
	Weak link	2.00	1.00	33%
	Moderate link	2.50	1.29	42%
	Close link	2.00	1.41	33% <sup>a</sup>
High School	No link	0	0	0
	Weak link	1.00	0.00	17%
	Moderate link	4.33	0.58	72%
	Close link	1.00	0.00	17%

<sup>a</sup> Total sums to greater than 100% because percentages are based on mean number of tasks and one rater only applied two (of four) rating options.

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

The extended standards should link to the full academic standards in performance expectations as well as content, although the depth of these expectations can be reduced for the alternate assessment. Several analyses were conducted to compare the performance levels specified in the Essence Statements to the full Minnesota Academic Standards. One analysis focused on the DOK ratings. Panelists worked together to achieve consensus DOK ratings on the Essence Statements and the benchmarks in the Minnesota Academic Standards separately. These ratings were analyzed for comparability.

Table 3.3 presents the percentage of Essence Statements per grade level rated as expecting performance at the same level, or higher or lower levels, as the full content standards. There is no minimum level of acceptable overlap in depth of knowledge based on the Links criteria; however, it is reasonable to expect that as many as half of the extended standards would require students to demonstrate performance at a lower level than the grade level content standards. Additionally, it would be problematic to find many (if any) extended standards with performance expectations at a higher level than the regular content standards.

**Table 3.3 Percentages of Essence Statements at Same, Lower, Higher Levels of Complexity Compared to Related Benchmarks.**

MTAS Grade Test	Percentage of Essence Statements at Varying Levels of Complexity		
	Same	Lower	Higher
5	50%	50%	0
8	50%	30%	16%
High School	67%	33%	0

For grades 5 and 8, panelists rated half of the Essence Statements (three of six) as assessing student knowledge at the same level of complexity as the benchmarks, while panelists for high school rated over half (four of six) statements as the same complexity level as the benchmarks. The grade 8 panelists rated one Essence Statement as higher in complexity than the related benchmarks. The table below (taken from the MTAS Test Specifications for Science, p.21) lists this Essence Statement and the corresponding benchmarks:

<b>Strand II – Physical Science</b>	Item Total for MCA-II	<b>MTAS Benchmark Number of Tasks</b>
	By Strand	
	10 – 12	
<b>Sub-strand B. Chemical Reactions</b> Standard: The student will differentiate between chemical and physical changes.	By Sub-strand	1
	0 – 2	
<b>Benchmark</b>	By Benchmark	
<b>6.II.B.1 and 6.II.B.2</b> <b>MCA-II Benchmark</b> The student will define chemical and physical changes. The student will observe that substances react chemically with other substances to form new substances with different characteristic properties. <b>MTAS Essence Statement</b> The student will recognize when matter changes to new substance(s). <b>MTAS Content Limits</b> Items may not use the term chemical change. Chemical changes are limited to changes commonly seen in daily life. Examples of chemical changes may include a burning match and rotting food.	0 – 2	

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

**Table 3.4 Mean Number of Essence Statements at Various Levels of Performance Centrality**

Grade	Performance Centrality Rating	Mean	SD	Percentage of Essence Statements per Rating
5	Same	0	0	0
	Partly similar	5.00	0.00	83%
	No similarity	1.00	0.00	17%
8	Same	1.75	0.50	29%
	Partly similar	3.75	1.26	63%
	No similarity	0.50 <sup>a</sup>	1.00	8%
High School	Same	0	0	0
	Partly similar	6.00	0.00	100%
	No similarity	0	0	0

<sup>a</sup> Only 1 panelist assigned this rating to an Essence Statement.

Grades 8 and high school passed this minimum level of acceptability with 92% (Same=29% and Partly Similar=63%) and 100% of the Essence Statements respectively. The grades 3-5 Essence Statements showed reasonable similarity in performance to the grade-level benchmarks, but failed to meet the minimum level of acceptability. Grade 5 in particular included one (which constitutes 17% of six statements) Essence Statement rated as entirely different compared to the targeted benchmark. For grade 8, a single panelist rated two Essence Statements as dissimilar to the grade-level benchmarks. Furthermore, the remaining three panelists rated four to five (of six) Essence Statements as ‘partly similar’ to the benchmarks. Based on these results, the Essence Statements for grade 8 and high school link sufficiently to the full content standards in performance expectations. However, since several Essence Statements for grades 5 and 8 were rated as different in performance than the benchmarks (even if for a small number of raters), Minnesota may wish to review these content expectations.

**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 grade levels. For this reason, the evaluation of content differentiation involves a comparison *between* grade-level content expectations. Panelists rated the Essence Statements between grade spans as to whether they evidenced broader, deeper, and newer knowledge, as well as if certain expectations represented prerequisite skills (see Appendix C for a more detailed explanation of the categories). Across these categories, panelists indicated whether the content differentiation of Essence Statements between grades was clear (C), partial (P), limited (L), or None (N). These ratings were conducted collaboratively among panelists to achieve consensus evaluations. According to the Links 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 Consensus Ratings on Content Differentiation between Grade Spans of Science MTAS Essence Statements**

Criterion	Grade 5	Grade 8	High School
Broader	P	L	N
Deeper	P	C	L
Prerequisite	L	L	C
New	P	C	C
Identical	N <sup>a</sup>	N	L

<sup>a</sup> 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, the degree of content differentiation varied across dimensions and grade levels. The grades 3-5 Essence Statements were rated as at least partially linked to the grade level content on 3 of 4 dimensions. For grades 8 and high school, panelists indicated that the content clearly increased from the previous grade span on a couple of dimensions; however, they also considered some content to show limited or no differences on other dimensions.

Generally, raters indicated the clearest content differentiation for the New dimension, where all panelists rated the level of content differentiation as clear or partial. At a minimum, panelists seemed to believe higher grade Essence Statements introduced knowledge or skills not previously addressed in lower grade Essence Statements. Ratings for the Broader dimension seem to indicate the least differentiation. Based on their comments, panelists seem to have indicated less differentiation in this dimension because of the diversity of the Essence Statements across grade level. That is, panelists noted Essence Statements for upper grades seemed to be more based on new content than extended (i.e., broader) applications of previously introduced content. This explanation is consistent with the high ratings of differentiation for the New dimension. For the Identical dimension, panelists for grades 5 and 8 indicated no identical content across Essence Statements; the high school panelists indicated limited

Identical standards, noting some overlap. Because the Essence Statements should demonstrate a progression across grade levels, ratings of “None” or “Limited” for the Identical dimension are actually positive, since they indicate more differentiation.

***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 Essence Statements by providing two separate ratings. First, we asked panelists to determine the level of communication required by each Essence Statement in order for students to demonstrate knowledge. The common categories applied, according to the Links method, include three ability levels for students with significant disabilities<sup>10</sup>:

- 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 can 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 is always ‘symbolic,’ it becomes much more difficult for modifications to be made and still retain comparability in content and performance at the more basic levels of communication. Instead, it is preferable that the access point of most extended standards (and assessment tasks) be pre-symbolic. Thus, the minimum level of acceptability is that the access point for at least 90% of the Essence Statements should be pre-symbolic.

Table 3.6 presents panelists’ mean ratings on the communication levels needed to demonstrate content knowledge for each set of grade span Essence Statements. The Essence Statements for grades 5 and high school clearly meet the minimum requirement for acceptability.

<sup>10</sup> In addition to rating descriptions in the Links 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>

**Table 3.6 Mean Number of Essence Statements Rated at Each Level of Symbolic Communication**

Grade	Level of Symbolic Communication Required	Mean	SD	Percentage of Essence Statements per Rating
5	Pre-symbolic	5.75	0.50	96%
	Early Symbolic	1.00	0.00	17%
	Full Symbolic	0	0	0
8	Pre-symbolic	4.33	2.89	72%
	Early Symbolic	5.50	0.71	92%
	Full Symbolic	0	0	0
High School	Pre-symbolic	6.00	0.00	100%
	Early Symbolic	0	0	0
	Full Symbolic	0	0	0

The results for grade 8 do not meet the minimum; however, some interpretation of the means is necessary in this case. The panelists essentially were split in their ratings between pre-symbolic and early symbolic for grade 8. For example, raters 1 and 4 judged most Essence Statements to be pre-symbolic, while raters 2 and 3 judged most statements to be early symbolic. This inconsistency may be partly attributable to disagreements among special educators (not just within this panel) concerning the distinction between these two categories and the behaviors that reflect each category. It is also possible that the Essence Statements for grade 8 need further clarification. Regardless, it is important to note that none of the six Essence Statements for grade 8 (or grades 5 and high school) received a rating of symbolic as the access point for students. Consequently, most of the content expectations should be accessible to a wide range of students.

The second 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 need for assistive technology demonstrate the knowledge expected by these Essence Statements? 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. Table 3.7 indicates the percentage of Essence Statements that were judged as accessible to all groups.

**Table 3.7 Mean Number of Essence Statements Rated as Accessible to All Students**

Grade	Mean	SD	Percentage of Essence Statements per Rating
5	4.50	1.00	75%
8	6.00	0.00	100%
High School	6.00	0.00	100%

All Essence Statements for grades 8 and high school were judged as accessible to a wide range of students. However, two Essence Statements from the grades 3-5 span raised concerns over accessibility for each panelist. Specifically, panelists indicated in their comments that visually impaired students may have difficulty effectively demonstrating knowledge, and that there even could be safety concerns, for the following two grade 3-5 science Essence Statements (MTAS Test Specifications for Science, pp.14-15):

<b>Strand I – History and Nature of Science</b>	<b>Item Total for MCA-II</b>	<b>MTAS Benchmark Number of Tasks</b>
	<b>By Strand</b>	
	9 – 11	
<b>Sub-strand B. Scientific Inquiry</b>	<b>By Sub-strand</b>	
Standard: The student will understand the nature of scientific investigations (3.I.B); the student will participate in a controlled scientific investigation (4.I.B); the student will understand the process of scientific investigations (5.I.B).	4 – 6	
<b>Benchmark</b>	<b>By Benchmark</b>	
<b>3.I.B.2 and 5.I.B.1</b>		
<b>MCA-II Benchmark</b> The student will participate in a scientific investigation using appropriate tools. The student will perform a controlled experiment using a specific step-by-step procedure and present conclusions supported by the evidence.	1 – 2	
<b>MTAS Essence Statement</b> The student will identify tools appropriate for a given scientific investigation. <b>MTAS Content Limits</b> Appropriate tools are limited to thermometers, hand lenses, rulers, and balances. Items may require students to choose a tool that is most appropriate to a particular task in a simple scientific investigation.		2

<b>Strand II – Physical Science</b>	Item Total for MCA-II	<b>MTAS Benchmark Number of Tasks</b>
	By Strand	
	8 – 10	
<b>Sub-strand A. Structure of Matter</b> Standard: The student will know that heating and cooling may cause changes to the properties of a substance.	By Sub-strand	2
	3 – 4	
Benchmark	By Benchmark	
<b>4.II.A.1</b> <b>MCA-II Benchmark</b> The student will observe that heating and cooling can cause changes in state. <b>MTAS Essence Statement</b> The student will identify changes in states of matter (solid, liquid, and gas). <b>MTAS Content Limits</b> Changes in state are limited to changes in water or other substances that are commonly seen in a student’s daily life. Examples may include ice melting (solid to liquid), water boiling (liquid to gas), and water freezing (liquid to solid). Changes are limited to one change in state per example.	1 – 2	

A review of these content expectations may be warranted to consider whether they are appropriate for all students who may take the MTAS assessment. If reasonable modifications can be made for these students to fairly demonstrate knowledge, it may be useful to provide examples of modifications in the Test Specifications for Science or other documents used by teachers.

### ***Inter-Rater Reliability***

Since panelists provided independent ratings for all of the analyses reported here, it is useful to consider how well these panelists were in agreement with each other. We used the intraclass correlation (ICC) statistic to measure the agreement between panelists’ ratings. This statistic indicates the amount of agreement by producing a statistic between 0 and 1. A positive correlation approaching 1 represents high agreement. Conversely, as the correlation approaches 0, or is negative, we interpret these outcomes to mean that panelists assigned quite different ratings to the same dimension resulting in weak agreement. Similar to Webb (2005), we applied the following decision criteria for judging the correlation outcomes:

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

We calculated correlations across panelists (n=4 per grade) on each of the rating dimensions: (a) age appropriateness, (b) content centrality, (c) performance centrality, (d) communication levels, and (e) accessibility. These results are presented in table 3.8.

**Table 3.8 Inter-Rater Agreement on Ratings for Essence Statements per Grade Level**

	Intraclass Correlation Coefficients		
	Grade 5	Grade 8	High School
Age Appropriate	0.89	1.00	1.00
Content Centrality	0.88	0.77	0.71
Performance Centrality	1.00	0.72	1.00
Communication Levels	0.87	1.00	1.00
Accessibility	0.97	1.00	1.00

Across grade levels, panelists agreed substantially in their ratings with good to exact agreement on all dimensions.

### ***Summary and Discussion of Essence Statements and Minnesota Academic Standards***

For this alignment evaluation, panelists reviewed the Essence Statements for science in two ways. First, they evaluated the content alignment (Criteria 2, 3, and 5 from the Links method) between the grade span Essence Statements and the corresponding Minnesota Academic Standards. Second, these panelists rated the accessibility and appropriateness (Criterion 7) of the content for this population of students. The results of this review indicated that these panelists found most Essence Statements across grade levels to link sufficiently with the Minnesota Academic Standards and to provide appropriate access to all types of students. In addition, no grade span Essence Statements stood out as particularly problematic across more than one dimension.

Table 3.9 displays the overall conclusions regarding content alignment between the Essence Statements and Minnesota Academic Standards for science. These judgments are based on whether the Essence Statements achieved acceptable levels of linkage with the full content standards for each set of grade span Essence Statements.

- 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%)

As with most alignment reviews, some areas of weakness were identified. These outcomes warrant further explanation. Specifically, while all Essence Statements were rated as linked to a Minnesota benchmark, some Essence Statements for grades 5 and 8 did not sufficiently link to the *central* grade level content.

Concerning content differentiation, most panelists agreed that the expectations of the Essence Statements do require students to demonstrate more advanced knowledge as grade levels increase. However, the high school grade span Essence Statements showed inconsistent increases in content expectations relative to earlier grades. In particular, panelists determined that these Essence Statements as a whole did not cover a broader range of content compared to lower grade expectations, and they considered some of the content to be too similar. Panelists did find that the Essence Statements for high school showed evidence of more complex knowledge expectations (deeper) and *some* new information not presented at earlier grades.

**Table 3.9 Summary Conclusions on Alignment of Essence Statements to Minnesota Academic Standards for Science on Links Criteria 2, 3, and 5**

Grade Level Tests	Criterion 2	Criterion 3		Criterion 5
	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	Partial	Partial	Partial
8	High	Partial	Partial	Partial
High School	High	High	High	Partial

Table 3.10 displays the overall conclusions on content accessibility pertaining to Performance Accuracy (content accessibility) for the Essence Statements. For this criterion, conclusions reflect overall judgments of acceptability based on access to the content expectations<sup>11</sup>.

<sup>11</sup> 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 3.10 Summary Conclusions on Performance Accuracy (Links Criterion 7) of Essence Statements for Science**

Criterion 7		
Grade Level Tests	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	Good	Acceptable
8	Acceptable	Excellent
High School	Excellent	Excellent

Panelists for the high school science assessment considered all of the Essence Statements to be appropriate for students at different communication and ability levels. The results for the grades 3-5 and grades 6-8 Essence Statements indicated show good to excellent accessibility; however, each set of content expectations included one or two statements judged by panelists to be limited in access for particular students. For example, panelists mostly agreed that students of all levels of communication can access the content and demonstrate knowledge for these Essence Statements. However, they felt that some types of students with particular disabilities may be disadvantaged by some of the content expectations. Specifically, panelists indicated in their comments that visually impaired students may have difficulty demonstrating knowledge effectively, and that there even could be safety concerns for these students in demonstrating one science Essence Statement. Thus, it is still appropriate to consider reviewing the performance expectations for these Essence Statements to ensure broad accessibility to all students.



## Chapter 4 Results: Science MTAS Tasks and Essence Statements

In this chapter, we report the results of panelists' ratings on the Science MTAS tasks per grade assessment. We present the results on the Links Criteria 1 through 7, followed by inter-rater agreement results.

### *Results on Science MTAS Tasks based on Links Criteria*

Ratings involved evaluation of the assessment relative to the Essence Statements on all seven of the Links criteria. As with Chapter 3 in which we compared the Essence Statements to the Minnesota Academic Standards, most results reflect mean ratings on a series of scales. Mean ratings were derived from frequency counts (per panelist) of how many Essence Statements were rated at each response option. From these counts, we then calculated the mean number of Essence Statements per response option (across panelists) for each rating scale. At least 90% of tasks must achieve acceptable ratings to demonstrate linkage to grade-level content for each Links criterion.

**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, math, 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 grade level science assessments as to whether each task does focus primarily on academics. Results of this analysis are presented in Table 4.1. At least 90% of tasks should be rated as academic.

**Table 4.1 Mean Number of Tasks Rated as Academic by Panelists**

Grade	Tasks Rated as Academic		
	Mean Number of Tasks	SD	Percentage of Tasks
5	12.75	0.96	85%
8	15.00	0.00	100%
High School	15.00	0.00	100%

Overall, results show that these panelists considered most tasks to be academic in nature. For grade 5 in particular, two panelists (of four) rated several tasks as focused more on functional skills without clear connection to the academic content of the Essence Statements.

**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 4.2 shows the mean number and percent 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.

**Table 4.2 Mean Percentage of Tasks at Various Levels of Age Appropriateness**

Grade	Performance Centrality Rating	Mean	SD	Percentage of Tasks per Rating
5	Adapted	10.00	4.69	67%
	Neutral	4.25	5.19	28%
	Inappropriate	1.00	0.00	7%
8	Adapted	15.00	0.00	100%
	Neutral	0	0	0
	Inappropriate	0	0	0
High School	Adapted	14.25	0.96	95%
	Neutral	1.50	0.71	10%
	Inappropriate	0	0	0

All grade-level science tests surpassed the minimum requirement. It is interesting to note that approximately four to five tasks for grade 5 were rated as neutral. While above the minimum when tasks rated as adapted and neutral are combined, it is surprising that, at least according to these panelists, the test includes several tasks that could be given at any grade level. Finally, grade 5 included one task rated as off-grade by all panelists.

**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 Essence Statements to determine the extent to which the tasks assess grade-level content. Several analyses were performed on these ratings. First, we reviewed the number of tasks that were linked to at least one Essence Statement. Table 4.3 shows that all raters considered all tasks for each grade science test to link to an Essence Statement.

**Table 4.3 Mean Number of Tasks Linked to Essence Statements**

Grade	Tasks Linked to Essence Statements		
	Mean Number of Tasks	SD	Percentage of Tasks
5	15.00	0.00	100%
8	15.00	0.00	100%
High School	15.00	0.00	100%

We also asked panelists to evaluate *how well* the tasks targeted the content expectations. At least 90% of tasks should be judged as moderate to closely linked with the Minnesota benchmarks for acceptability. Table 4.4 presents the mean number and percent of tasks that fell into each category based on panelists' ratings.

**Table 4.4 Mean Percent of Items at Various Levels of Content Centrality**

Grade	Content Centrality Rating	Mean	SD	Percentage of Items per Rating
5	No link	0.25 <sup>a</sup>	0.50	2%
	Weak link	2.75	1.71	18%
	Moderate link	0.50 <sup>a</sup>	1.00	3%
	Close link	8.50	2.65	57%
8	No link	0.25 <sup>a</sup>	0.50	2%
	Weak link	2.33	1.53	16%
	Moderate link	6.67	2.89	44%
	Close link	8.00	2.45	53%
High School	No link	0.50 <sup>a</sup>	1.00	3%
	Weak link	4.25	1.50	28%
	Moderate link	3.75	1.50	25%
	Close link	6.50	1.73	43%

<sup>a</sup> Only 1 panelist assigned this rating to a performance task.

Grade 8 was the only test that met the minimum level of acceptability (97% of tasks rated as moderate to closely linked). Across panelists for grade 8, a mean of 6.67 tasks were rated as moderately linked and a mean of 8 tasks were rated as closely linked. For each grade assessment, two to three panelists rated several tasks as weakly linked to the targeted benchmarks in the Minnesota Academic Standards. In addition, one person per grade level panel rated at least one task as not linked to the Essence Statements.

***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 Essence Statements should require students to make some type of distinction. Table 4.5 shows the mean number of tasks rated as retaining all (same performance), some, or none of the performance expectations of the corresponding benchmarks. At least 90% of tasks should receive ratings of ‘Some’ or ‘All.’

**Table 4.5 Mean Percent of Tasks at Various Levels of Performance Centrality**

Grade	Performance Centrality Rating	Mean	SD	Percentage of Tasks per Rating
5	All	8.00	3.61	53%
	Some	8.25	3.86	55%
	None	0.75 <sup>a</sup>	1.50	1%
8	All	7.75	4.99	52%
	Some	7.00	4.55	47%
	None	0.25 <sup>a</sup>	0.50	2%
High School	All	5.50	2.12	37%
	Some	9.00	2.94	60%
	None	4.33	4.04	29%

<sup>a</sup> Only 1 panelist assigned this rating to any performance task.

All grade level tests met the minimum level of acceptability (90%) of tasks assessing students on at least some of the same performance expectations as the benchmarks. However, it is noteworthy that several tasks in each grade assessment were rated as assessing entirely different performance expectations. Minnesota may wish to review these tasks to determine if greater linkage with the grade level benchmarks could be attained.

**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.*

Criterion 4 incorporates the Webb alignment statistics. For each alignment indicator, we present the mean results of panelists’ ratings for each grade test. Results are reported at the strand level. Thus, the mean ratings reported indicate which content strands associated with the Essence Statements are covered well on the assessment, based on panelists’ evaluations.

**Categorical Concurrence.** In the previous section on Content Centrality under Criterion 3, we presented results on whether or not, and how well, each task matched to content expectations. For this analysis, we focus on the content expectations to determine *which* Essence Statements were assessed. Categorical concurrence describes the extent to which the Essence Statements are covered by the assessment. For a regular assessment, Webb recommends a minimum of six test questions assessing each content strand to adequately cover that content; but for an alternate

assessment, the criterion is one performance task per standard. This change is appropriate because alternate assessments tend to include considerably fewer items compared with a regular assessment. In addition, a single task may assess multiple content expectations<sup>12</sup>.

Table 4.6 summarizes the MTAS alignment results for categorical concurrence. As Table 4.6 indicates, all three grade tests met the Webb alignment criterion of at least one task per strand for all Essence Statements. Moreover, the distribution of items across strands matches closely with the distribution outlined in the Minnesota test specifications.

**Table 4.6 Summary of Categorical Concurrence Results for Science MTAS by Grade Level**

Grade Level	Mean Number of Items per Strand				Strands with at Least One Task
	History and Nature of Science	Physical Science	Earth and Space Science	Life Science	
5	2	2	2	3	4 of 4
8	2	2	2	3	4 of 4
High School	1	NA <sup>a</sup>	NA	8	2 of 2

<sup>a</sup> NA = Strand not taught or assessed at this grade level.

**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. To make these judgments, reviewers first determined the DOK level for each Essence Statement of each strand using a rating scale (see Appendix C for the LINKS 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. Table 4.7 shows the mean percentage of tasks rated at each DOK level per grade level.

<sup>12</sup> The psychometric trade-off is that fewer items per strand may lead to a decrease in scoring accuracy.

**Table 4.7 Mean Percentage of Items at Various Levels of DOK**

Grade	Item DOK Rating	Mean	SD	Percentage of Items per Rating
5	None	0	0	0
	Attention	0	0	0
	Memorize/recall	5.75	3.86	38%
	Performance	2.00	0.00	13%
	Comprehension	8.25	3.95	55%
	Application	0	0	0
	Analysis, Synthesis, Evaluation	0	0	0
8	None	0	0	0
	Attention	0	0	0
	Memorize/recall	5.00	0.00	33%
	Performance	2.33	0.58	16%
	Comprehension	8.25	5.38	55%
	Application	3.33	0.58	22%
	Analysis, Synthesis, Evaluation	5.00	0.00	33%
High School	None	1.50	0.71	10%
	Attention	0	0	0
	Memorize/recall	6.33	1.15	42%
	Performance	2.33	2.31	16%
	Comprehension	1.67	0.58	11%
	Application	4.50	4.95	30%
	Analysis, Synthesis, Evaluation	5.67	1.53	38%

We then compared these two separate judgments about cognitive complexity (one for the Essence Statement, one for the task) to determine the proportion of tasks written at the appropriate level. Webb refers to this comparison as *depth-of-knowledge consistency*.

Table 4.8 summarizes the depth-of-knowledge consistency results for each grade level of the Science MTAS assessment. Since reviewers evaluated depth of knowledge at the most specific level of the standards document (Essence Statements), the table refers to consistency between the performance tasks and the Essence Statements to which they were matched. Results are summarized in terms of the percentage of tasks with cognitive complexity ratings at or above the rating for the corresponding Essence Statement. Webb's suggested criterion for this alignment indicator is the same as for a regular assessment – at least 50% of the tasks should have complexity ratings at or above the level of the corresponding Essence Statement per strand.

**Table 4.8 Summary of Depth-of-Knowledge Results for Science MTAS by Grade Level**

Grade Level	Percentage of Tasks with DOK At or Above the Level of the Benchmarks per Strand				Number of Strands Assessed Adequately	Specific Strands Assessed Inadequately
	History and Nature of Science	Physical Science	Earth and Space Science	Life Science		
5	51	100	38	91	3 of 4	Earth and Space Science
8	8	100	88	75	3 of 4	History and Nature of Science
High School	75	NA <sup>a</sup>	NA	61	2 of 2	None

<sup>a</sup> NA = Strand not taught or assessed at this grade level.

As Table 4.8 indicates, most strands met Webb's target of 50% at most grade levels. The two exceptions are the Earth and Space Science strand for grade 5 and the History of Nature and Science strand for grade 8. Because the MTAS is administered only to students with the most severe cognitive disabilities, Minnesota may wish to review the appropriateness of the cognitive benchmarks expected of these students. If the level of the content standards is appropriate, more complex items may need to be developed for the grade 5 Earth and Space Science and grade 8 History and Nature of Science items, where more of the strands failed to demonstrate adequate DOK consistency. Conversely, Minnesota may wish to explore whether its Physical Science items are too complex or whether its Physical Science Essence Statements are not complex enough, as all of the performance tasks are written at or above the level of complexity of the Essence Statements.

**Range of Knowledge.** Range of knowledge measures how fully the tasks cover each of the Essence Statements within each strand. The assessed Essence Statements within a strand should be linked with at least one performance task. Webb's minimum level of acceptability for range-of-knowledge correspondence is 50% per strand. This means that at least 50% of the Essence Statements must be matched to one or more tasks.

Table 4.9 summarizes the range-of-knowledge results for each grade level of the MTAS. We computed the number of Essence Statements covered for each strand separately for each panelist and then averaged across panelists to obtain the summary alignment indicator. As Table 4.9 demonstrates, all strands were covered adequately across all grade levels. That is, at least 50% of the Essence Statements were linked to at least one performance task for all strands at all grade levels.

**Table 4.9 Summary of Range-of-Knowledge Results for Science MTAS by Grade Level**

Grade Level	Percent of Essence Statements per Strand Matched to at Least One Task				Number of Strands Assessed Adequately	Specific Strands Assessed Inadequately
	History and Nature of Science	Physical Science	Earth and Space Science	Life Science		
5	100	100	100	100	4 of 4	None
8	100	88	100	100	4 of 4	None
High School	100	NA <sup>a</sup>	NA	90	2 of 2	None

<sup>a</sup> NA = Strand not taught or assessed at this grade level.

**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 objective per standard. The number of tasks should be distributed rather evenly between the objectives for each standard to achieve good balance.

The content balance is determined by calculating an index, or score, for each standard<sup>13</sup>. According to Webb, the minimum acceptable index for a single standard is 70 (on a scale of 0 to 100, with 100 representing perfect balance). To be clear, a standard may include more objectives than reviewers actually linked to performance tasks. Thus, only those objectives actually used by the reviewers are included in calculations of the balance index.

Table 4.10 summarizes the results on balance of content representation per grade level of the Science MTAS. As the table demonstrates, all content strands met Webb's criterion of a balance index of at least 70 across all grade levels. In fact, the balance index was a perfect 100 for all strands except Life Science, indicating very strong distribution of content across Essence Statements.

<sup>13</sup> 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>.

**Table 4.10 Summary of Balance-of-Knowledge Representation Results for Science MTAS by Grade Level**

Grade Level	Balance Index per Strand				Strands with Adequate Balance	Strands with Limited Balance
	History and Nature of Science	Physical Science	Earth and Space Science	Life Science		
5	100	100	100	83	4 of 4	None
8	100	100	100	83	4 of 4	None
High School	100	NA <sup>a</sup>	NA	81	2 of 2	None

<sup>a</sup> NA = Strand not taught or assessed at this grade level.

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

As with the evaluation of the Essence Statements, 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 MTAS performance tasks between grade level tests (higher and lower). Table 4.11 shows panelists' consensus ratings across the various dimensions using the rating scheme of clear (C), partial (P), limited (L), or none (N). Each test should evidence at least partially different content per dimension relative to higher or lower grade tests.

**Table 4.11 Consensus Ratings on Content Differentiation between Grade Level Science MTAS Assessments**

Criterion	Grade 5	Grade 8	High School
Broader	C	L	C
Deeper	C	C	C
Prerequisite	P	L	C
New	C	L	C
Identical	L	L	L

Across all grades, panelists noted some differentiation in content, including ratings of clear content differentiation on the Deeper and Broader dimensions. These results indicate deeper mastery is required for higher grade skills or knowledge. For the Identical dimension, the ratings of "None" or "Limited" are reversed because here they imply limited or no identical content in performance tasks across grades. Panelists determined that the presence of identical content between grade level assessments was limited, which corresponds with their judgments of deeper and broader knowledge assessment across grades. Finally, a review of panelists' comments lends further support to conclusions of appropriate increases in depth and breadth of knowledge at higher grade levels (see Appendix B).

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

The sixth Links criterion pertains to demonstration of student learning. Thus, this criterion focuses more on accessibility to students than on content alignment. The alternate assessment should allow students with disabilities to demonstrate academic skills or knowledge acquired from their coursework on the assessment. To determine the extent to which the MTAS *enables* students to demonstrate this learning, panelists evaluated the scoring rubric and achievement level descriptors relative to the assessment. Panelists worked together for consensus to determine whether the assessment allowed for demonstration of high, low, or no evidence of student learning. These 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 4.12 includes the group consensus ratings on the degree of student inference evident in the Science MTAS assessment per grade level. Results for this criterion were quite mixed across grade levels. Panelists determined that it is possible to make strong student inferences about learning, but only based on certain dimensions for each grade test. Grade 8 in particular received a rating of 'high inference' on only one dimension (Level of Accuracy), and panelists could not find any evidence for being able to evaluate new learning or show generalizability in knowledge. The findings on

Program Quality Indicators, however, merely reflect the fact that MTAS scoring is not based on these factors so this dimension is not applicable (which is appropriate based on the Links method).

**Table 4.12 Degree of Inference Evident on Student Learning in Science MTAS Assessments**

Dimensions	Grade 5	Grade 8	High School
Level of Accuracy	H <sup>a</sup>	H	L
Level of Independence	H	L	H
New Learning	L <sup>b</sup>	N	H
Generalization across People and Settings	N <sup>c</sup>	N	H
Generalization across Materials and Activities	L	L	L
Standard Setting	L	L	N/A
Program Quality Indicators	N/A <sup>d</sup>	N/A	N/A

<sup>a</sup> H = high student inference

<sup>b</sup> L = low student inference

<sup>c</sup> N = no student inference

<sup>d</sup> N/A = not applicable; the MTAS assessment does not include Program Quality indicators.

Based on a review of panelists' written comments, it is possible that there may have been some confusion among panelists over certain dimensions. For example, the high school panelists chose to provide an "N/A" rating for the Standard Setting dimension because they felt that they could not make an accurate rating since cut scores were not available to them. We still encourage Minnesota to consider these outcomes due to the importance of accurate measurement of student learning.

**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.

Criterion 7 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 Essence Statements, panelists rated tasks on the levels of communication required to respond and the access to each type of student who takes the assessment. In addition, panelists evaluated each task on whether modifications or supports can be made for different types of students without substantially altering the target content.

Table 4.13 gives mean ratings on the communication levels required of students in order to respond to the science tasks. At least 90% of tasks should be rated as pre-symbolic for reasonable access by all students. As the table indicates, only high school met this minimum level. Three panelists rated all tasks as pre-symbolic, while the fourth panelist rated all tasks to be early symbolic. Grade 5 also evidences reasonable access

across tasks because three panelists rated all tasks as pre-symbolic, while the fourth panelist rated 40% (n=6) of tasks as pre-symbolic and the remaining tasks (n=9) as symbolic. Thus, the majority of panelists considered the access point for the grade 5 tasks to be pre-symbolic. For grade 8, the results were rather mixed, with three of four panelists rating the majority of tasks as early symbolic.

**Table 4.13 Mean Percentage of Tasks at Various Levels of Symbolic Communication**

Grade	Level of Communication Rating Categories	Mean	SD	Percentage of Tasks per Rating
5	Pre-symbolic	12.75	4.50	85%
	Early Symbolic	0	0	0
	Symbolic	2.25 <sup>a</sup>	4.50	15%
8	Pre-symbolic	8.00	9.90	53%
	Early Symbolic	14.00	1.00	93%
	Symbolic	0.50 <sup>a</sup>	1.00	3%
High School	Pre-symbolic	15.00	0.00	100%
	Early Symbolic	3.75 <sup>a</sup>	7.5	25%
	Symbolic	0	0	0

<sup>a</sup> Only 1 panelist assigned this rating to a performance task.

Concerning the accessibility of task content to students with a variety of disabilities, all panelists for grades 8 and high school considered all tasks to be accessible to a wide range of students, as shown in Table 4.14. In contrast, the grade 5 panelists considered just over half of these tasks to have broad accessibility across students. As a result, these tasks should undergo a more comprehensive bias review with special education experts to determine if the current tasks can be modified for better accessibility, or if new tasks must be implemented.

**Table 4.14 Mean Numbers of Tasks Rated Accessible to Students**

Grade Level	Mean	SD	Percentage of Tasks Accessible
5	9.25	3.59	62%
8	15	0	100%
High School	15	0	100%

Finally, panelists evaluated the Science MTAS tasks on an additional dimension under Criterion 7 not included for the Essence Statements. A common approach to administering an alternate assessment is for teachers to make modifications in test structure or offer supports (i.e., assistive technology; prompts if needed) as appropriate for a given student. Panelists were asked to rate each task as to whether they could in fact be modified or supports offered, particularly without altering the target of the assessment. Table 4.15 includes the mean number of tasks panelists found amenable to these types of changes.

**Table 4.15 Mean Number of Tasks Amenable to Modifications or Supports**

Grade Level	Mean	SD	Percentage of Modifiable Tasks
5	11.25	2.06	75%
8	15.00	0.00	100%
High School	15.00	0.00	100%

As with several other rating dimensions, the grades 8 and high school panelists considered all tasks amenable to modifications or supports. However, the grade 5 panelists found several tasks that they felt may not be easily changed for all students. Unfortunately, panelists often were split on which tasks fell into this category. Such an outcome seems to suggest that these panelists struggled with some tasks to find appropriate modifications, which provides further support (in addition to findings on Communication Levels and Accessibility) for the need to review grade 5 science tasks.

### **Reliability Results**

In this section, we report on two types of agreement analyses on panelists' ratings. First, we indicate the inter-rater agreement levels between panelists on the ratings they assigned to tasks for various rating scales. Second, we compare panelists' ratings on content match to the test contractor's intended content match.

#### **Inter-Rater Reliability.**

We used the intraclass correlation (ICC) statistic to measure the agreement between panelists' ratings on the science performance tasks. This statistic indicates the amount of agreement by producing a statistic between 0 and 1. A positive correlation approaching 1 represents high agreement. Conversely, as the correlation approaches 0, or is negative, we interpret these outcomes to mean that panelists assigned quite different ratings to the same dimension resulting in weak agreement. Similar to Webb (2005), we applied the following decision criteria for judging the correlation outcomes:

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

We calculated correlations across panelists (n=4 per grade) on each of the rating dimensions used by panelists to evaluate the tasks: (a) age appropriateness, (b) content centrality, (c) performance centrality, (d) categorical concurrence, (e) DOK match, (f) communication levels, and (g) accessibility. These results are presented in Table 4.16.

**Table 4.16 Inter-Rater Agreement on Panelists' Ratings of Science MTAS Tasks per Grade Level**

	Intraclass Correlation Coefficients		
	Grade 5	Grade 8	High School
Age Appropriate	0.78	1.00	0.93
Content Centrality	0.72	0.30	0.65
Performance Centrality	0.74	0.56	0.81
Categorical Concurrence	0.98	0.93	0.92
Depth of Knowledge	0.51	0.30	0.79
Communication Levels	0.85	1.00	0.94
Accessibility	0.82	1.00	1.00

From Table 4.16, it is clear that panelists agreed strongly on some ratings, while showing distinct disagreement on other rating scales for each grade level test. Overall, panelists for the high school assessment showed the greatest agreement in assigned ratings across dimensions. Panelists for grade 8 seemed to be the most inconsistent in their ratings across the dimensions with several cases where they achieved perfect agreement (or 1.0) and other cases where raters rarely agreed across the 9 performance tasks.

### Panelist-Test Developer Analyses.

In addition to examining inter-rater agreement, we assessed the agreement between our panelists' judgments of assessed content and the Pearson item specifications for each performance task. Such a comparison provides an independent evaluation of the State's content assignment. Table 4.17 includes these comparisons by noting the percent of tasks with exact agreement between panelists and Pearson on content match.

**Table 4.17 Percentage Agreement between Panelists and Pearson on Assessment Target for Science MTAS Tasks**

Grade	Number of Tasks	Number of Raters	Percentage of Tasks with Exact Agreement
5	9	4	100%
8	9	4	100%
High School	9	4	73%

***Summary and Discussion of Science MTAS Tasks and Essence Statements***

Table 4.18 displays the overall conclusions regarding content alignment between the Science MTAS assessments and the Essence Statements. These judgments are based on whether the Essence Statements achieved acceptable levels of linkage with the full content standards for each grade test. The minimum level for each of the criteria in Table 4.18 is 90%.

- High linkage           - most 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 4.18 Summary Conclusions on Alignment of Science MTAS Assessments to Essence Statements for Links Criteria 1, 2, 3, 4, and 5**

	Criterion 1	Criterion 2	Criterion 3		Criterion 4		Criterion 5
Grade Level Tests	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 Essence Statements?	Does the performance of task link to expectations of the Essence Statements?	Do the tasks assess students at the appropriate breadth of knowledge? <sup>a</sup>	Do the tasks assess students at the appropriate depth of knowledge? <sup>b</sup>	Do the assessments show appropriate increases between grade levels?
5	Partial	High	Partial	High	High	Partial	Partial
8	High	High	High	High	High	Partial	Partial
High School	High	High	Partial	High	High	High	High

<sup>a</sup> This conclusion is based on a summary judgment across the Webb statistics of Categorical Concurrence, Range of Knowledge, and Balance of Knowledge. It is still important to consider each of the criteria separately as well.

<sup>b</sup> This conclusion is based on the results from the DOK consistency analyses.

Table 4.19 includes results relative to Criterion 7 of the Links 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 alignment.

- 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 4.19 Summary Conclusions on Accessibility (Links Criteria 6 and 7) of Science MTAS Assessments**

	Criterion 6	Criterion 7		
Grade Level Tests	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	Acceptable	Questionable	Acceptable
8	Questionable	Questionable	Excellent	Excellent
High School	Acceptable	Excellent	Excellent	Excellent



## Chapter 5 Results: Science MTAS Tasks and Alternate Achievement Standards

In this chapter, we describe the review of the Science MTAS assessments relative to the alternate achievement standards. This review involved an evaluation of performance alignment. Alternate achievement standards allow for classification of students into various performance categories based on their test scores.

Through a standard-setting process, states determine which scores on the assessment represent various levels of achievement by establishing a cut-off location, or “cut score,” between adjoining categories. As part of the standard-setting process, content and special education experts examine test items and use their professional judgment to define categories based on test performance. For all but strictly normative test results, some judgment is required of these experts to define exactly what test performance means. This is especially true of tests that categorize students into value-laden categories, such as ‘Proficient,’ for all of the NCLB assessments. For Minnesota, the standard-setting process resulted in four distinct achievement levels linked to cut scores: (a) Exceeds the Standards, (b) Meets the Standards, (c) Partially Meets the Standards, or (d) Does Not Meet the Standards.

The standard-setting process used by Minnesota relies on a “ordered item book” procedure in which judges review assessment items arrayed in a booklet by their relatively difficulty. Judges identify those locations in the booklet which seem to best distinguish between the expected performance levels for the four different Minnesota achievement levels. This use of actual items almost guarantees that the difficulty levels of the assessment will match the difficulty of the achievement levels. Nevertheless, a quality assessment should contain a set of items whose difficulties are arrayed across the range of the achievement levels. For example, an assessment with all “easy” items will not measure well in the top achievement levels.

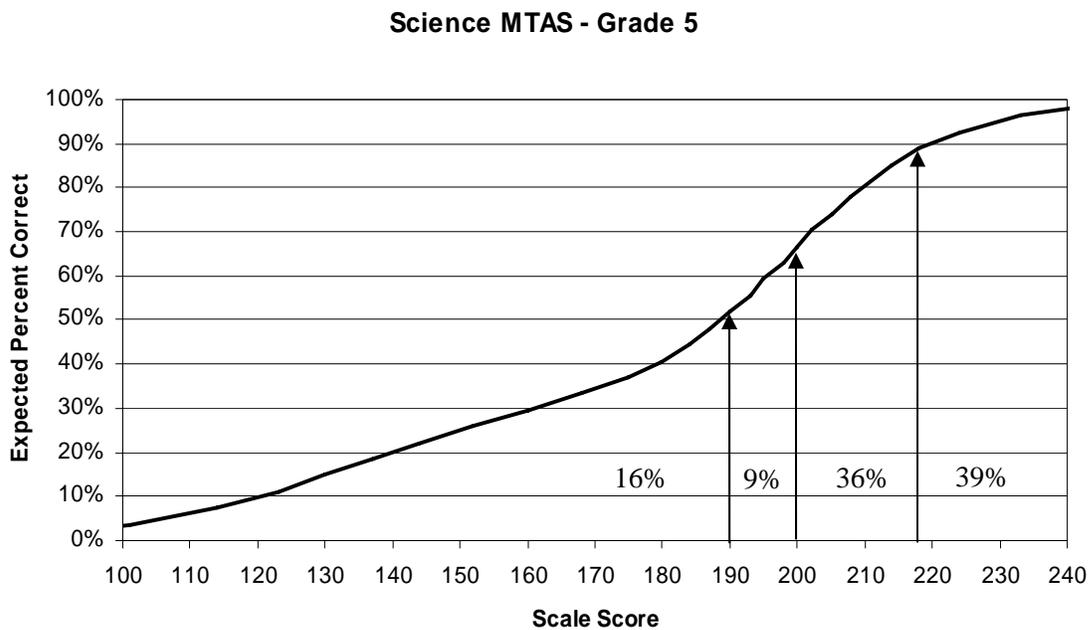
A convenient and informative vehicle for reviewing the MTAS assessments in relation to the achievement levels is the so-called “test characteristics curve” (TCC). Like other Minnesota assessments, the psychometrics for MTAS are based on Item Response Theory (IRT) which establishes a relationship between subject matter achievement and item performance. Thus, in each of the figures below, expected test performance, expressed as percentage of items correct, is shown as a function of achievement, expressed in the MTAS reporting scale. The higher a student’s achievement, the greater the percentage of test points attained. The critical issue for Chapter 5 is the location of the achievement level cut scores with regard to this relationship.

Note that in each of the figures below, the relationship is a curved, not straight, line. This is characteristic inherent in IRT. The curvilinear relationship tends to be flatter at the lower and higher levels of achievement and steeper in middle. An assessment functions best in the range of achievement where the curve slopes more steeply upward. Ideally, the assessment should also function best in the range of the

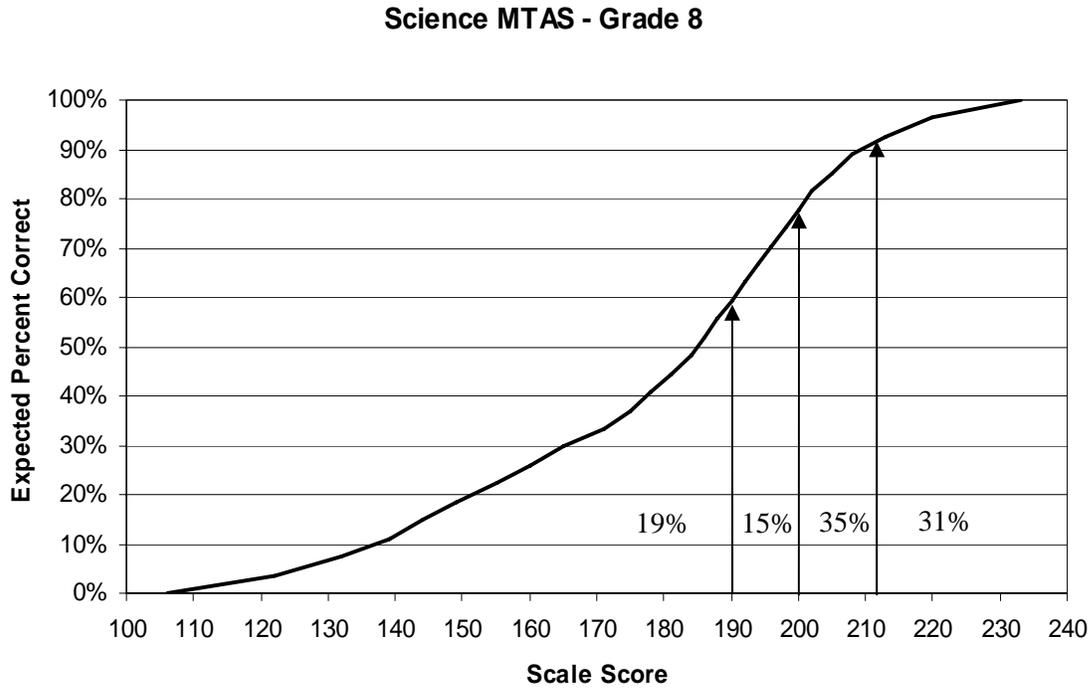
achievement level cut points. That is, the steeper parts of the TCC should cover the area of the achievement level cuts.

A lesser concern is that the majority of students score within the range at which the test functions well (and hopefully in the range containing the cut scores as described above). The assessment should be functioning in the range where most of the student population scores, assuming that most students score near the achievement levels. A convenient method for making this assessment is to examine the percentages of students within each achievement level. These percentages are noted in each of the following figures.

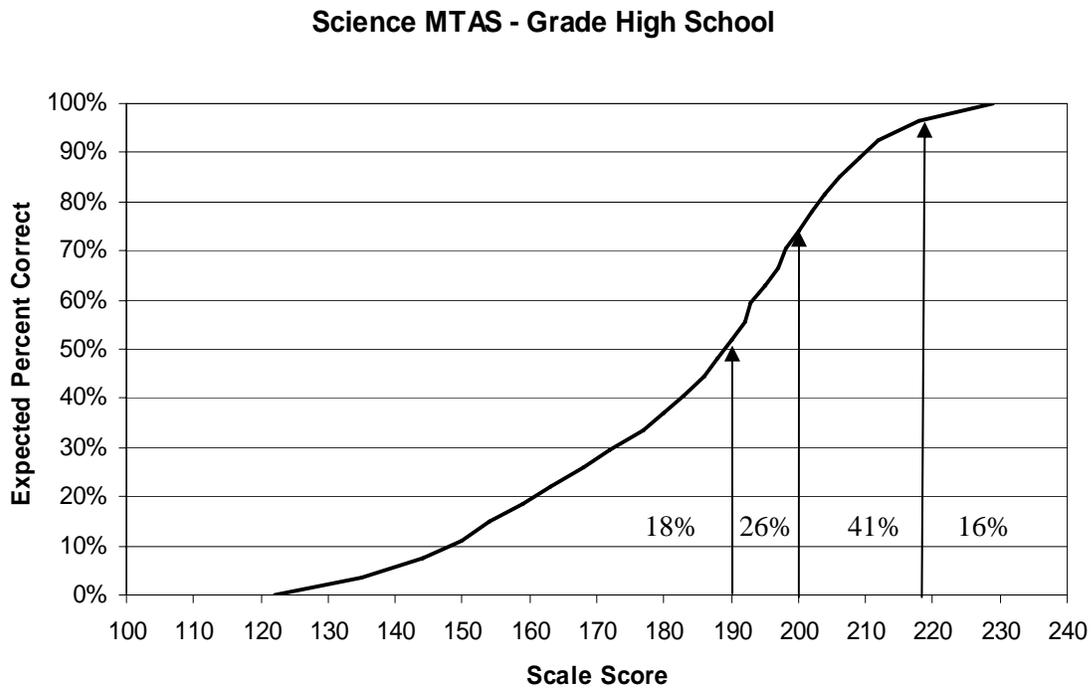
Please note that the TCC and student percentages in the figures are based on 2008 assessment forms and 2008 student results. Because items vary from year to year, the TCC will vary as well. One of the constraints of test construction, however, is that items be selected to produce similar curves across years. Student performance is expected to improve from year to year.



**Figure 5-1. Alignment of achievement levels for Grade 5 Science MTAS.**



**Figure 5-2. Alignment of achievement levels for Grade 8 Science MTAS.**



**Figure 5-3. Alignment of achievement levels for High School Science MTAS.**

### ***Summary and Discussion of Science MTAS Tasks and Alternate Achievement Standards***

The figures above indicate that the Science MTAS assessments generally function best in the range of the achievement level cut points. That is, in each case, the steeper parts of the TCCs cover the range of the achievement level cuts although the TCCs begins to flatten out, indicating less discrimination, in the location of the Exceeds/Meets cut. In terms of Minnesota's AYP scoring, discrimination between these top two categories is less important than discrimination among the other three categories. Increasing the overall difficulty of the assessment by replacing some of the easier tasks with more difficult tasks may improve the "meets" versus "exceeds" discrimination; however, if not done carefully, it could reduce the assessment's ability to discriminate among the three lower categories that are used to determine schools' AYP scores. The MTAS assessments must cover a large range of student abilities, which creates an equally large challenge. Therefore, this report will stop short of recommending a change in task difficulty based on these results. There are simply too many other considerations. As a result, we conclude that the Science MTAS assessments meet requirements of alignment of test difficulty with the achievement level standards, particularly for those categories that are used to determine schools' AYP status.

## Chapter 6 Summary and Recommendations

HumRRO conducted an alignment review of the Science MTAS assessment to evaluate the content alignment, as well as content accessibility, of the extended standards, alternate achievement standards, and alternate assessments. Alignment to the state academic content standards is a requirement of the No Child Left Behind Act of 2001, although alternate standards and assessments may be reduced in breadth and depth. Furthermore, all aspects of the assessment system must be accessible to the student for whom it was designed.

Three types of alignment evaluations were applied to the grade 5, 8, and high school Science MTAS tests: (a) alignment of the Essence Statements to Minnesota Academic Benchmarks for science, (b) alignment of the Science MTAS assessments to the Essence Statements, and (c) alignment of the Science MTAS assessments to the alternate achievement standards. The cumulative results point to reasonable content linkage of the assessment, Essence Statements, and alternate achievement standards with the content standards. Concerning accessibility, the content expectations and assessments were rated as accessible to a range of students with various disabilities. However, some features of the assessments and Essence Statements exhibited better content access and capacity for students to demonstrate knowledge than others per Links criterion and grade level. This conclusion includes the evaluation of the assessment to alternate achievement standards in terms of appropriately classifying students.

As with most reviews of state assessment systems, these findings suggest areas where Minnesota could strengthen the content alignment between the MTAS components as well as student access to this content. For this reason, HumRRO makes the following recommendations to Minnesota per assessment component. These recommendations focus on the more critical findings, including those portions of Tables 3.9 and 3.10 in Chapter 3 and Tables 4.18 and 4.19 in Chapter 4 highlighted in red (weak or questionable). However, some findings highlighted in yellow (partial or acceptable) in these tables also are included if the Essence Statements or assessment fall short on a serious issue, such as accessibility.

### Essence Statements for Science

- (1) ***Review the content differentiation between grade span Essence Statements.*** Each set of grade-span content expectations demonstrated some need for greater differentiation in content at increasingly higher grade levels. While panelists found broadening and deepening of knowledge expectations, increases were very limited in some cases. Minnesota may find this recommendation daunting because it would seem to require re-writing (and, hence, re-approval by the State) the Essence Statements. However, it may be that the current Essence Statements can be better differentiated by adding to the content limitations and/or including examples of how students might demonstrate knowledge

differently at higher grade levels. Some states include examples of performance activities for each content expectation per grade level

- (2) **Review the access points for each of the grade-span Essence Statements.** None of the Essence Statements were rated as highly exclusive – most were at least acceptable if not quite good in allowing student access and demonstration of knowledge. However, since student access is such a critical issue, we suggest that Minnesota re-examine the Essence Statements for grades 5 and 8 in particular. Such a task may involve additional bias reviews, or, as noted above, further explanation (content limitations, examples) of how teachers and test administrators might make these content expectations more appropriate within the MTAS Test Specifications for Science document.

### **MTAS Performance Tasks**

It is important to note that no panelists considered any of the tasks to display serious flaws that would warrant complete replacement of tasks. Instead, ratings and comments by panelists point to issues that could be improved upon for better student access.

- (1) **Review some performance tasks for the Grade 5 and high school assessments for clarity in targeted content (content centrality).** While panelists agreed with the State on the target of assessment in most cases, panelists also indicated that some (approximately 3 to 4) performance tasks did not always measure this content well. Ratings on these tasks also correspond with lower ratings on capacity for demonstrating achievement accessibility in many cases. These combined outcomes suggest that certain performance tasks may require additional review by content and special education experts. Based on panelists' comments, such a review may only require edits to task presentation or response card options, as opposed to a complete revision of the topic of the performance task or target of assessment.
- (2) **Improve the ability of the Science MTAS assessments to accurately demonstrate student knowledge.** This recommendation reflects the combined outcomes from student inference on achievement and performance accuracy based on accessibility. For the grades 5 and 8 assessments in particular, panelists found that some tasks do not allow for clear inference about student learning, which they partly attributed to limitations in access to certain student groups. One issue in particular noted by panelists is the fact that the Science MTAS assessment does not include a pretest, or baseline, which raised some concerns because the science assessments cover multiple grade content.

A second concern among panelists focused on the option for test administrators to apply a wide range of alternate materials for

modification, thus potentially reducing standardization. We agree with the latter point in part, but with the recognition that alternate assessments should allow for reasonable modification. Such comments by panelists seem to reflect concern that test administrators or teachers may not be well versed, or comfortable, with what counts as appropriate modifications. More direction as part of training on test administration may be appropriate.

## 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](http://www.naacpartners.org/LAL/documents/NAAC_AlignmentManualVer8_3.pdf)
- Minnesota Department of Education (January, 2008). *Minnesota test of academic skills (MTAS): Test specifications for science*. Roseville, MN: Minnesota Department of Education. Retrieved from: [http://education.state.mn.us/MDE/Accountability\\_Programs/Assessment\\_and\\_Testing/Assessments/MTAS/MTAS\\_Test\\_Specifications/index.html](http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Test_Specifications/index.html)
- 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 Minnesota, National Center on Educational Outcomes.
- 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 science education*. Washington, D.C.: Council of Chief State Schools Officers.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states. (Research Monograph 18)*. Madison, WI: National Institute for Science Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852)

## Appendix A Webb Alignment Results per Grade Level Assessment

### *Webb Alignment Results*

The following tables include complete statistical results on the Webb alignment indicators (Links Criterion 4: Content Coverage).

#### **Categorical Concurrence**

The categorical concurrence results for grades five, eight, and high school of the Science MTAS assessment are presented below. Each table includes: the target number of items from the test blueprint; the mean number of items matched by panelists; the standard deviation among panelists' ratings; and, the final alignment conclusion (Yes or No). The bottom row indicates the percentage of standards that met the minimum alignment criterion.

***Table A- 1. Categorical Concurrence for Science MTAS, Grade 5: Mean Number of Performance Tasks per Strand***

Title of Strand	Number of Tasks per Strand			At Least One Task per Strand
	Target # Items from Blueprint	Mean Tasks Matched	Standard Deviation	
History and Nature of Science	2	1.75	0.50	Y
Physical Science	2	2.25	0.50	Y
Earth and Space Science	2	2.00	0.00	Y
Life Science	3	3.00	0.00	Y
Total	9	9.00		
Percent of strands with at least one task				100%

***Table A- 2. Categorical Concurrence for Science MTAS, Grade 8: Mean Number of Performance Tasks per Strand***

Title of Strand	Number of Tasks per Strand			At Least One Task per Strand
	Target # Items from Blueprint	Mean Tasks Matched	Standard Deviation	
History and Nature of Science	2	2.25	0.50	Y
Physical Science	2	2.00	0.00	Y
Earth and Space Science	2	1.75	0.50	Y
Life Science	3	3.00	0.00	Y
Total	9	9.00		
Percent of strands with at least one task				100%

**Table A- 3. Categorical Concurrence for Science MTAS, High School: Mean Number of Performance Tasks per Strand**

Title of Strand	Number of Tasks per Strand			At Least One Task per Strand
	Target # Items from Blueprint	Mean Tasks Matched	Standard Deviation	
History and Nature of Science	2	1.25	0.50	Y
Physical Science	0	--	--	--
Earth and Space Science	0	--	--	--
Life Science	7	7.75	0.50	Y
Total	9	9.00		
Percent of strands with at least one task				100%

**Depth-of-Knowledge Consistency**

The Depth-of-Knowledge (DOK) consistency results for grades five, eight, and high school of the Science MTAS assessment are presented below. The tables present the results from the comparison between the depth-of-knowledge expected in the standards and the depth-of-knowledge 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 level. Standards with at least 50% of items at the same (or above) DOK level met the minimum criterion.

**Table A- 4. Depth-of-Knowledge Consistency for Science MTAS, Grade 5: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives**

Title of Strand	Mean Tasks per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
History and Nature of Science	1.75	50	57.74	13	25.00	38	47.87	Y
Physical Science	2.25	0	0.00	0	0.00	100	0.00	Y
Earth and Space Science	2.00	63	25.00	0	0.00	38	25.00	N
Life Science	3.00	8	16.67	83	19.25	8	16.67	Y
Percent of strands with 50% of item DOK at or above objective DOK:								75%

**Table A- 5. Depth-of-Knowledge Consistency for Science MTAS, Grade 8: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives**

Title of Strand	Mean Tasks per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
History and Nature of Science	2.25	92	16.67	8	16.67	0	0.00	N
Physical Science	2.00	0	0.00	38	25.00	63	25.00	Y
Earth and Space Science	1.75	13	25.00	63	47.87	25	50.00	Y
Life Science	3.00	25	31.91	8	16.67	67	27.22	Y

Percent of strands with 50% of item DOK at or above objective DOK: 75%

**Table A- 6. Depth-of-Knowledge Consistency for Science MTAS, High School: Mean Percent of Performance Tasks with DOK Below, At, and Above DOK Level of Objectives**

Title of Strand	Mean Tasks per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
History and Nature of Science	1.25	25	50.00	75	50.00	0	0.00	Y
Physical Science	--	--	--	--	--	--	--	--
Earth and Space Science	--	--	--	--	--	--	--	--
Life Science	7.75	38	16.91	52	35.22	9	18.75	Y

Percent of strands with 50% of item DOK at or above objective DOK: 100%

### Range-of-Knowledge Correspondence

The results for Range-of-Knowledge correspondence for grades five, eight, and high school of the Science MTAS assessment are presented below. The tables include the mean number, standard deviation, and percentage of Essence Statements by strand. A minimum of 50% of Essence Statements within each strand should be matched to at least one item.

**Table A- 7. Range-of-Knowledge for Science MTAS, Grade 5: Mean Percent of Essence Statements per Strand Linked with Performance Tasks**

Title of Strand	Number of Essence Statements	Mean Tasks per Strand	Range of Essence Statements			Range-of-Knowledge Target Met
			Essence Statements with At Least One Task		% of Total Essence Statements per Strand	
			M	S.D.	M	
History and Nature of Science	1	1.75	1.00	0.00	100	Y
Physical Science	1	2.25	1.00	0.00	100	Y
Earth and Space Science	2	2.00	2.00	0.00	100	Y
Life Science	2	3.00	2.00	0.00	100	Y
Percentage of strands with 50% of essence statements linked to at least one item						100%

**Table A- 8. Range-of-Knowledge for Science MTAS, Grade 8: Mean Percent of Essence Statements per Strand Linked with Performance Tasks**

Title of Strand	Number of Essence Statements	Mean Tasks per Strand	Range of Essence Statements			Range-of-Knowledge Target Met
			Essence Statements with At Least One Task		% of Total Essence Statements per Strand	
			M	S.D.	M	
History and Nature of Science	1	2.25	1.00	0.50	100	Y
Physical Science	2	2.00	1.75	0.00	88	Y
Earth and Space Science	1	1.75	1.00	0.00	100	Y
Life Science	2	3.00	2.00	0.00	100	Y
Percentage of strands with 50% of essence statements linked to at least one item						100%

**Table A- 9. Range-of-Knowledge for Science MTAS, High School: Mean Percent of Essence Statements per Strand Linked with Performance Tasks**

Title of Strand	Number of Essence Statements	Mean Tasks per Strand	Range of Essence Statements			Range-of-Knowledge Target Met
			Essence Statements with At Least One Task		% of Total Essence Statements per Strand	
			M	S.D.	M	
History and Nature of Science	1	1.25	1.00	0.50	100	Y
Physical Science	0	--	--	--	--	--
Earth and Space Science	0	--	--	--	--	--
Life Science	5	7.75	4.50	0.50	90	Y
Total	6	9.00				
Percentage of strands with 50% of essence statements linked to at least one item						100%

**Balance-of-Knowledge Representation**

The results for Balance-of-Knowledge representation for grades five, eight, and high school of the Science MTAS assessment are presented below. The tables also include the percentage of items linked to each strand. The minimum acceptable balance index is 70 out of 100.

**Table A- 10. Balance-of-Knowledge Representation for Science MTAS, Grade 5: Mean Balance Index per Strand**

Title of Strand	Balance-of-Knowledge Representation						Balance Index Target Met
	Essence Statements per Strand	Mean Essence Statements Linked with Tasks	Mean Tasks per Strand	Mean % of Tasks (of total) Linked to Strand	Mean Balance Index	S.D.	
	M	M	M	M	S.D.		
History and Nature of Science	1	1.00	1.75	19	100	0.00	Y
Physical Science	1	1.00	2.25	25	100	0.00	Y
Earth and Space Science	2	2.00	2.00	22	100	0.00	Y
Life Science	2	2.00	3.00	33	83	0.00	Y
Percentage of standards with a balance of representation index of 70 or greater							100%

**Table A- 11. Balance-of-Knowledge Representation for Science MTAS, Grade 8: Mean Balance Index per Strand**

Title of Strand	Balance-of-Knowledge Representation						
	Essence Statements per Strand	Mean Essence Statements Linked with Tasks	Mean Tasks per Strand	Mean % of Tasks (of total) Linked to Strand	Mean Balance Index		Balance Index Target Met
		M	M	M	M	S.D.	
History and Nature of Science	1	1.00	2.25	26	100	7.45	Y
Physical Science	2	1.75	2.00	20	100	0.00	Y
Earth and Space Science	1	1.00	1.75	20	100	0.00	Y
Life Science	2	2.00	3.00	34	83	0.00	Y
Percentage of standards with a balance of representation index of 70 or greater							100%

**Table A- 12. Balance-of-Knowledge Representation for Science MTAS, High School: Mean Balance Index per Strand**

Title of Strand	Balance-of-Knowledge Representation						
	Essence Statements per Strand	Mean Essence Statements Linked with Tasks	Mean Tasks per Strand	Mean % of Tasks (of total) Linked to Strand	Mean Balance Index		Balance Index Target Met
		M	M	M	M	S.D.	
History and Nature of Science	1	1.00	1.25	14	100	0.00	Y
Physical Science	0	--	--	--	--	--	--
Earth and Space Science	0	--	--	--	--	--	--
Life Science	5	4.50	7.75	86	81	5.59	Y
Percentage of standards with a balance of representation index of 70 or greater							100%

## Appendix B

### Summary of Panelist Comments on Essence Statements and Performance Tasks

The tables below summarize panelists' feedback on the performance tasks for the Science MTAS assessment. Comments were reported only if more than one panelist made the same observation. To maintain test security, individual item identifiers are not presented, nor are any comments that would reveal the content of a task. Additionally, comments have been stripped of any wording that would reveal test content and arranged into more general categories or comment types.

***Table B-1. Summary of Repeated Panelist Comments on Science MTAS Essence Statements***

Grade	Comment	Percent of panelists	Number of panelists with comment
5	Items would be difficult to adapt for visually impaired [comment repeated for multiple items]	100%	4
8	Difficult to determine best Essence Statement match	50%	2
High School	Inappropriate/Implausible Item Distractors	75%	3
	Connection between task and Essence Statement seems forced	75%	3

***Table B-2. Summary of Repeated Panelist Comments on Science MTAS Performance Tasks***

Grade	Comment	Percent of panelists	Number of panelists with comment
5	Mismatch between Essence Statement and benchmark	100%	4
8	Essence Statement omits portions of the benchmark critical to understanding	75%	3



## Appendix C

### Sample Alignment Review Materials

*Panelists received the following instruction sheet as a reference guide corresponding with verbal instructions from HumRRO facilitators.*

### Science MTAS Panelist Task Instructions

	Rating Task	Documents Needed	File Format
1	DOK of Minnesota Academic Standards for Science	(1) Minnesota Academic Standards for Science (HumRRO Coded) (2) MTAS Code Descriptions	Print Copy Print copy
2	MN Essence Statements for Science	(1) Minnesota Essence Statements for Science (HumRRO Coded) (2) MTAS Code Descriptions (3) MTAS_EssenceRatings_Science	Print copy Print copy Excel spreadsheet
3	Individual MTAS tasks	(1) Minnesota Essence Statements for Science (HumRRO Coded) (2) MTAS Code Descriptions (3) MTAS_ItemRatingForm_Science (4) MTAS Test Documents a. Test Administration Manual (includes Tasks and Scoring Rubric) b. Response Cards c. Presentation Pages (online version available for review)	Print copy Print copy Excel spreadsheet Print copies
*4	'Whole Test' Barriers	(1) MTAS_WholeTestRatings_Science (2) MTAS Code Descriptions	Excel spreadsheet Print copy
*5	Scoring criteria	(1) MTAS_ScoringRatings_Science (2) MTAS Test Administration Manual (3) Alternate Achievement Standards (4) MTAS Code Descriptions (see Scoring Inferences)	Excel spreadsheet Paper copy Paper copy Print copy
*6	Content differentiation across grades	(1) MTAS_ContentDiff_TestRatings_Science (2) MTAS_ContentDiff_EssncRatings_Science (3) Essence Statements for grade spans (4) MTAS Code Descriptions	Excel spreadsheet Excel spreadsheet Print copies Print copy

\* These rating tasks will be performed if there is time.

## 1 Rate DOK of Minnesota Academic Standards

Using the 'Minnesota Academic Standards-Science' printouts, assign a depth-of-knowledge rating to each benchmark of the Minnesota Academic Standard. You may simply write down your DOK ratings next to each benchmark and HumRRO Code. First, you will rate the benchmarks independently. Then, we will come to consensus on the ratings (3/4 majority). The consensus ratings will be retained for analysis.

## 2 Rate the MN Essence Statements

Open the Excel file 'MTAS\_EssenceRatings\_Science'. Click on the worksheet for your grade span. Evaluate the Essence Statements on all of the dimensions (columns) in the form. We will rate the Essence Statements on each dimension independently. Group consensus will be obtained on DOK ratings. We will discuss discrepant ratings on other dimensions if there is time.

- A. Assign a depth-of-knowledge rating to each Essence Statement using the DOK Code List.
- B. Determine whether the standards/benchmarks listed are the best match for the Essence Statement by indicating 'Y' (yes) or 'N' (no). If the Essence Statement matches a different standard/benchmark better, please use the 'Minnesota Academic Standards-Science' sheets to find a code for the alternate benchmark you chose and enter in Notes/Comments.

For the following ratings, use the codes listed on the 'MTAS Code Descriptions' handout.

- C. Indicate *how well* you think that the Essence Statement actually links to listed benchmark. Please reserve the use of a code of '1' (No Link) for two circumstances: (1) the Essence Statement does not link to any standard/benchmark, (2) the Essence Statement does not link to the standard/benchmark listed (but does link to alternate standard you chose above).
- D. Determine to what extent the Essence Statement measures student performance expected in the standard/benchmark. NOTE: If you chose an alternate standard/benchmark, evaluate the Essence Statement against that standard instead of the one listed.
- E. Evaluate whether the Essence Statement is appropriate for the chronological age at which the content is measured. Content may be grade-appropriate, off-grade level, or grade-neutral (meaning that the content/topic could be assessed at any grade).
- F. Evaluate the level of symbolic communication required to demonstrate content knowledge. 'Symbolic communication' can include use of pictures, symbols, signs, and speech. NOTE: Please consider the lowest functioning student who could access this task.
- G. Evaluate the accessibility of the Essence Statement content expectations for various disability groups. If the statement is accessible, enter a 'Y' (yes). If you think that the content is NOT accessible by some groups, enter 'N' (no) and provide an annotation in the Notes/Comments column to indicate those groups negatively affected.

## 3 Rate individual MTAS tasks

Open the Excel 'MTAS\_ItemRatingForm\_Science' file. Click on the appropriate grade spreadsheet. You will be rating each individual task (15 total) independently. We will discuss discrepant ratings if there is time.

***Only a few panelists may have time to complete the following ratings, depending on time.***

**4 Rate ‘Whole Test’ barriers to demonstrating student knowledge**

Open the Excel ‘MTAS\_WholeTestRatings\_Science’ file. Click on the appropriate grade worksheet. Make an evaluation of the test as a whole on the dimensions listed. 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.

**5 Rate scoring criteria (including scoring rubric and achievement descriptors)**

Open the ‘MTAS\_ScoringRatings\_Science’ spreadsheet and click on the appropriate grade spreadsheet. Rate the scoring rubric and achievement standards (found in the Test Manual) on the extent to which they allow for the demonstration of student learning. These documents should provide information about student performance rather than system or teacher performance. Refer to the MTAS Code Descriptions sheet for explanation of codes.

**6 Rate content differentiation across grades**

This task involves two sets of ratings. Open the ‘MTAS\_ContentDiff\_EssncRatings\_Science’ worksheet. Click on the appropriate grade span. Provide a holistic judgment about the differences found across the grade levels using the MTAS Code Descriptions sheet. In addition, document evidence that supports your ratings by providing brief descriptions.

Perform the same ratings on the tests using the MTAS\_ContentDiff\_TestRatings\_Science’ spreadsheet.

Panelists received the following coding sheet as a reference guide to each rating scale.

### MTAS Code Descriptions

#### Depth of Knowledge (DOK) (for Academic Standards, Essence Statements, and MTAS tasks)

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 Accessibility Dimensions (for use with Essence Statements and MTAS Tasks)

Category	Code	Description
<b>Academic</b>	A	Academic
	F	Foundational
	N	Neither foundational or academic
<b>Standard Match</b>		See Minnesota Extended Standards for MTAS printouts (8.5 x 14 paper)
<b>Content Centrality</b>	1	No link
	2	Weak link
	3	Moderate link
	4	Close link
<b>Age Appropriate</b>	A	Adapted from grade-level content
	I	Inappropriate; off-grade content
	N	Neutral; content is not age-bound
<b>Performance Centrality</b>	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
<b>Barriers to Demonstrating Knowledge</b>		
Symbolic Communication	A	Awareness/Pre-symbolic (gesture, purposeful moving toward object)
	E	Early Symbolic
	S	Symbolic (pictures, symbols, signs, speech)
Accessibility	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).
<b>Modifications/Supports</b>	Y	Yes, modifications and supports can be provided for this item.
	N	No, this item is not amenable to supports or modifications.

## Scoring Inferences

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	

### Content Differentiation across Grades

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

*Panelists received the Minnesota Academic Standards for science coded for data entry into rating forms. The content of the standards was extracted exactly from the full Minnesota Academic Standards document. Only a portion of the coded standards is replicated below grade 3 as an example.*

Grade Level	Strand	Sub-Strand	Standard	Benchmarks	HumRRO Code
GRADE 3	I. HISTORY AND NATURE OF SCIENCE	A. Scientific World View	The student will understand the use of science as a tool to examine the natural world.	1. The student will explore the use of science as a tool that can help investigate and answer questions about the environment.	31111
GRADE 3	I. HISTORY AND NATURE OF SCIENCE	B. Scientific Inquiry	The student will understand the nature of scientific investigations.	1. The student will ask questions about the natural world that can be investigated scientifically.	31211
				2. The student will participate in a scientific investigation using appropriate tools.	31212
				3. The student will know that scientists use different kinds of investigations depending on the questions they are trying to answer.	31213

Panelists received the Essence Statements coded for data entry into rating forms. The content of the Essence Statements was extracted exactly from the Test Specifications for Science. Only a portion of the coded Essence Statements is replicated below for grades 3-5 as an example.

Essence Statements	HumRRO Codes
<b>Strand I – History and Nature of Science</b>	
<b>Sub-strand B. Scientific Inquiry</b> Standard: The student will understand the nature of scientific investigations (3.I.B); the student will participate in a controlled scientific investigation (4.I.B); the student will understand the process of scientific investigations (5.I.B).	
<b>Benchmark</b>	
<b>3.I.B.2 and 5.I.B.1</b> <b>MCA-II Benchmark</b> The student will participate in a scientific investigation using appropriate tools. The student will perform a controlled experiment using a specific step-by-step procedure and present conclusions supported by the evidence.	
<b>MTAS Essence Statement</b> The student will identify tools appropriate for a given scientific investigation.	<b>31212</b> <b>51211</b>
<b>MTAS Content Limits</b> Appropriate tools are limited to thermometers, hand lenses, rulers, and balances. Items may require students to choose a tool that is most appropriate to a particular task in a simple scientific investigation.	

Panelists reviewed the Essence Statements using the following rating form in electronic format. The format of the rating form was identical for each grade span.

Evaluation of Science Essence Statements: Grades 3-5								
Topic	DOK of Essence Statement	Standard Match	Content Centrality	Performance Centrality	Age Appropriate	Barriers to Demonstrating Knowledge		Notes/Comments
Rating Code Options	Which DOK level does Extended Standard assess?  See coding sheet	Does the Extended Standard link to the academic standard listed?  Y, No=see codes	How well is content linked to academic standards?  1- No link Weak link Moderate link link 2- 3- 4- Close link	Does Extended Standard measure performance of Academic Standard?  N, S, A	Is Extended Standard grade-level appropriate?  A, N, I	What level of symbolic communication does Extended Standard require?  A, E, S	Is Extended Standard accessible to different disability groups?  Y, N	If you provide a low rating or 'No' answer to any dimensions, please explain your rating below.
Essence Statements								
1								
2								
3								
4								
5								
6								

Panelists reviewed the individual Science MTAS performance tasks using the following rating form in electronic format. The format of the rating form was identical for each grade span.

Evaluation of MTAS Tasks: Grade 5										
Topic	Academic	Standard Match	Content Centrality	Item DOK	Performance Centrality	Age Appropriate	Barriers to Demonstrating Knowledge			Notes/Comments
Rating Code Options	N, F, A	See standard codes	How well is task content linked to standards? 1- No link Weak link 3- Moderate link 4- Close link	Which DOK level does task assess? 2- Levels 0, 1, 2, 3, 4, 5, 6 (See DOK Scale)	Does task measure performance of standard? N, S, A	Is task content based on grade-level content? A, N, I	What level of symbolic communication does task require? N, S, E	Is task accessible to different disability groups? Y, N	Can task be modified/supports provided without changing meaning or difficulty? Y, N	If you provide a low rating or 'No' answer to any dimensions, please explain your rating below.
Task Number										
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Panelists reviewed and compared the Essence Statements for science and the Science MTAS assessments for grades 5, 8, and high school on Criterion 5: Content Differentiation using the following rating form in electronic format. The format of the rating form was identical for the Essence Statements and the Science MTAS assessments.

Content Differentiation Across Grade-Level Essence Statements		
Vertical Relationships	Content across Grade Levels	List Evidence to Support your Ratings
Rating Code Options	C, P, L, N	
Broader		
Deeper		
Prerequisite		
New		
Identical		

Panelists reviewed the Science MTAS assessments for grades 5, 8, and high school on Criterion 6: Achievement using the following rating form in electronic format. The format of the rating form was identical for each grade assessment.

<i>Student Learning Evident from Scoring Procedures</i>		
	<b>Evidence of Student Learning</b>	<b>Rationale for Rating (indicate evidence)</b>
Rating Code Options	H, L, N	
Level of Accuracy		
Level of Independence		
New Learning		
Generalization across people and settings		
Generalization across materials and activities		
Standard Setting		
Program Quality Indicators		

Panelists reviewed each Science MTAS assessment as a whole for Criterion 7: Performance Accuracy (Potential Barriers) using the following rating form in electronic format. The format of the rating form was identical for each grade span.

Barriers to Demonstrating Student Knowledge

Considerations		Type of Student								
Please enter Y or N in each cell to indicate 'yes' or 'no'.		Visually impaired/legally blind	Hearing impaired	Deaf/blind	Nonverbal - printed words	Nonverbal - pictures	Nonverbal - manual signs	Nonverbal - eye gaze	Verbal but no use of hands	Communicates with objects or by indicating yes/no
1	Provision for students with these characteristics?									
2	Student can do AA as designed with flexibility built into tasks?									
3	Student can do AA with accommodations (no change to meaning)?									
4	Student can do with modifications/supports (may change meaning)?									

Please enter Y or N in each cell to indicate 'yes' or 'no'.	
5	Can the assessment capture responses for students without clear, intentional communication (even at nonsymbolic level)?
6	Are accommodations, modifications, and supports defined sufficiently to maintain standardized administration?