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. Specifically, MDE wanted an evaluation of the alignment between the Science MTAS grade-level assessment, the extended benchmarks[4], and the Minnesota Academic Standards for Science[5]. 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, which occurred June 19-20, 2012.

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. 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 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" (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 more fairly and accurately demonstrate the achievement 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 for

[4] Extended standards can be found in the MTAS Test Specifications: http://education.state.mn.us/MDE/EdExc/Testing/TestSpec/

[5] Minnesota Academic Standards can be found at http://education.state.mn.us/MDE/EdExc/StanCurri/K-12AcademicStandards/

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

Professional Position	Number of Panelists	Mean Years of Experience	Special Certifications	Region of Origin in Minnesota			Gender		Ethnicity		
				7-County Metro	Greater Minnesota	MPLS/ St Paul	M	F	White, Non-Hispanic	Hispanic	Black, Non-Hispanic
Grade 5											
Teacher	3	19.33	2	0	3	0	0	3	3	0	0
Administrator	1	25.00		1	0	0	1	0	1	0	0
Grade 8											
Teacher	3	25.33	5	2	0	1	0	3	3	0	0
Administrator	1	14.00	2	1	0	0	0	1	0	0	1
High School											
Teacher	4	31.25	10	2	2	0	0	4	4	0	0
Administrator											

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

Test Forms. Reviewers evaluated the 2012 Science MTAS test form per grade-span. Table 2.2 below lists the number of operational performance tasks and field test tasks per test. In addition, we list the number of extended benchmarks per grade level assessed by each test.

Table 2.2. Characteristics of the Science MTAS Tests Reviewed

Grade Level	Number of Operational Tasks	Number of Field-Test Tasks	Number of Extended Benchmarks
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 LAL 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 with instruction sheets enumerating the alignment tasks that they needed to complete as well as code sheets listing the depth-of-knowledge ratings and other possible ratings for each task (see Appendix C).

Procedures. HumRRO conducted this alignment review on June 19-20, 2012. 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.

Following the general introduction, panelists began working within their content groups. Three groups of four panelists reviewed each grade-span test (i.e., grades 3-5, grades 6-8). HumRRO staff supervised the groups.

Within their small groups, HumRRO staff further trained reviewers using sample benchmarks and assessment tasks. Regarding instructions on how to rate benchmarks and tasks, 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 benchmarks or tasks because reviewers were valued as content experts. Each panelist received a laptop computer 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 DOK level of each grade-level extended benchmark. 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 extended benchmark. A volunteer scribe within each group recorded these consensus ratings.

Next, reviewers rated the extended benchmarks on a variety of factors, including (a) whether the Minnesota Academic Standards benchmark listed is the best match, (b) how well the extended benchmark links to the Minnesota Academic Standards benchmark, (c) whether the extended benchmark measures student performance of the Minnesota Academic Standards benchmark, (d) whether the extended benchmark 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 extended benchmark is accessible to various disability groups. These ratings were made individually; no consensus ratings were obtained.

Reviewers then received more specific instructions for rating performance tasks. For training, HumRRO staff facilitated reviewers in evaluating and discussing sample tasks as a group. After completing sample tasks, reviewers individually rated performance tasks on electronic rating forms on their computers. The panelists rated the tasks on a variety of factors, including (a) whether the extended benchmark listed is the best match, (b) how well the task links to the extended benchmark, (c) whether the task measures student performance of the extended benchmark, (d) whether the task 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 task is accessible to various disability groups. In addition, reviewers were instructed to assign a *primary extended benchmark* to a task based on a judgment that a task clearly measured this extended benchmark. Furthermore, reviewers could assign an *additional extended benchmark* only if the task seemed to assess another extended benchmark as clearly as the primary extended benchmark. 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. In High School, one panelist's data was not captured for rating performance tasks because the data was accidentally overwritten by the panelist. The panelist contributed in every other way to the group and was not anomalous for the data that was collected.

Finally, panelists worked in their small groups to develop consensus ratings for two 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 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. Typically, reviewers develop consensus ratings of the extent to which content differs across grades to assess the LAL Criterion 5: Content Differentiation. This criterion was not assessed for the Science MTAS because the test itself is a grade-span test.

Chapter 3: Results: Extended Benchmarks 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 extended benchmarks compared to the Minnesota Academic Standards for Science. These analyses relate to Criteria 2, 3, and 7 of the LAL method.

Results on Extended Benchmarks based on LAL Criteria

Panelists rated the extended benchmarks 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 extended benchmarks were rated at each response option per panelist for each scale. For most scales, we then calculated the mean number of extended benchmarks per response option (across panelists) from the frequency counts. Finally, we determined the percentage of extended benchmarks rated at each level per rating scale based on the means. Results of these analyses are presented for each set of extended benchmarks per grade-span. Each grade-span includes six extended benchmarks.

We point to several features of the results for more accurate interpretation. First, since the calculation of the percentage of extended benchmarks is based on the mean ratings, the total percentages across a rating scale per grade may sum to above 100%. Second, it is important to keep in mind that these percentages are based on four panelists' ratings of six extended benchmarks. In other words, a small number of raters evaluated a small number of extended benchmarks. Finally, most LAL criteria include a minimum number of extended benchmarks (generally 90%) needed to demonstrate reasonable linkage with the full content standards.

> ***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 extended benchmarks. For this evaluation, panelists were asked whether the content of the Science extended benchmarks is appropriate for the age and grade level indicated. Several response options were possible:

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

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 extended benchmarks receiving that rating across panelists. Column 5 represents this same mean as a percentage of the total number of extended benchmarks per grade. For this criterion, at least 90% of extended benchmarks should be rated as 'adapted' or 'neutral'[6].

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

Table 3.1. Mean Number of Extended Benchmarks Rated as Age Appropriate

Grade	Age-Related Content	Mean	SD	Percentage of Extended Benchmarks per Rating[a]
5	Adapted	3.50	1.29	58%
	Neutral	1.25	0.50	21%
	Inappropriate	0.50	0.58	8%
8	Adapted	5.75	0.50	96%
	Neutral	0.00	0.00	0%
	Inappropriate	0.00	0.00	0%
High School	Adapted	4.25	0.50	71%
	Neutral	0.50	0.58	8%
	Inappropriate	1.25	0.50	21%

[a] Total may sum to above 100% because percentages are based on mean numbers.

Eight percent of the extended benchmarks in grade 5 and twenty-one percent in High School were judged inappropriate by raters. This means that at least one extended benchmark in grade 5 and High School was rated by panelists as containing content that is off-grade level. There was one extended benchmark, dealing with water collection, in grade 5 for which three of the four panelists did not provide a rating. The one panelist who did rate this benchmark gave it an acceptable rating. All six extended benchmarks were rated as clearly adapted from appropriate grade-level content in grade 8.

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 extended benchmarks 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 extended benchmarks 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 extended benchmark linked to the benchmark.

Concerning overall content match, panelists at each grade level rated most of the Science extended benchmarks as matched to a primary benchmark. Tables 3.2 to 3.4 show the primary benchmarks assigned to each extended benchmark by panelists. The number of panelists that gave the match is presented as well as any comments regarding the degree of match.

Table 3.2. Grade 5 Science Extended Benchmarks Matched to Primary Benchmarks

Extended Benchmark	Benchmark	Number of Raters	Comments
3.1.3.4.1	3.1.3.4.1	4	
4.2.1.2.2	4.2.1.2.2	3	
4.3.2.3.1	4.3.2.3.1	1	
5.3.4.1.3	5.3.4.1.3	4	
3.4.1.1.2	3.4.1.1.2	4	
4.4.4.2.1	4.4.4.2.1	4	

Table 3.3. Grade 8 Extended Benchmarks Matched to Primary Benchmarks

Extended Benchmark	Benchmark	Number of Raters	Comments
6.1.2.1.1	6.1.2.1.1	4	
8.2.1.2.2	8.2.1.2.2	4	
6.2.2.2.1	6.2.2.2.1	4	
8.3.1.2.1	8.3.1.2.1	4	
7.4.1.1.2	7.4.1.1.2	3	
7.4.4.2.1	7.4.4.2.1	3	

Table 3.4. High School Extended Benchmarks Matched to Primary Benchmarks

Extended Benchmark	Benchmark	Number of Raters	Comments
9.1.1.2.1	9.1.1.2.1	4	Two different expectations of students – recognizing a scientific experiment is at a far different level than making a hypothesis or making a conclusion.
9.4.1.1.2	9.4.1.1.2	4	Structure/function is very different than the concept of homeostasis. Also doesn't deal with body systems.
9.4.2.1.2	9.4.2.1.2	3	Expectation is below grade level. Benchmark asks HOW ecosystem will change. Simply recognizing that it will change for the extended benchmark is a different expectation.
9.4.3.2.1	9.4.3.2.1	3	This extended benchmark does not seem to match any High School benchmark.
9.4.3.3.5	9.4.3.3.5	4	Seems like the extended benchmark has lost the definitions needed to understand the expectations.
9.4.4.1.2	9.4.4.1.2	4	

For the second evaluation, panelists reviewed each grade-span extended benchmark for the degree of link to the central content targeted by the benchmarks. In this case, panelists used the following four-point scale to determine how well the extended benchmark 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 3.5 shows that only grade 8 extended benchmarks surpassed this minimum. Panelists rated only 58% and 51% of the extended benchmarks in grade 5 and High School, respectively, to link sufficiently ('moderate' or 'close' link) with the benchmarks. In both of these grades, there were 8% or roughly one of the extended benchmarks that were rated as entirely different from the full benchmarks. In grade 5 and High School, it may be necessary to review the extended benchmarks. Table 3.6 lists the extended benchmarks that were rated, on average, as having 'no' or 'weak' link. See Tables 3.2 and 3.4 for additional information regarding comments on extended benchmarks.

Table 3.5. Mean Number of Extended Benchmarks at Various Levels of Content Centrality

Grade	Content Centrality Rating	Mean	SD	Percentage of Extended Benchmarks per Rating[a]
5	No link	0.50	0.58	8%
	Weak link	1.25	0.50	21%
	Moderate link	1.50	1.00	25%
	Close link	2.00	0.00	33%
8	No link	0.00	0.00	0%
	Weak link	0.00	0.00	0%
	Moderate link	0.50	1.00	8%
	Close link	5.25	1.50	88%
High School	No link	0.50	0.58	8%
	Weak link	2.25	0.50	38%
	Moderate link	2.25	0.50	38%
	Close link	0.75	0.96	13%

[a] Total may sum to above 100% because percentages are based on mean numbers.

Table 3.6. Extended Benchmarks Rated Less Than Moderate Link

Grade	Extended Benchmark	Average Content Centrality Rating	Number of Panelists
5	3.1.3.4.1	2.25	4
	4.3.2.3.1	0.75	1[a]
	4.4.4.2.1	1.50	4
High School	9.4.1.1.2	2.00	4
	9.4.2.1.2	2.25	3
	9.4.3.2.1	1.75	4

[a] Only one panelist provided a benchmark and content centrality rating for this extended benchmark.

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

The extended benchmarks 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 extended benchmarks to the full Minnesota Academic Standards. One analysis focused on the

DOK ratings. Panelists worked together to achieve consensus DOK ratings on the extended benchmarks and the benchmarks in the Minnesota Academic Standards separately. These ratings were analyzed for comparability.

We compared these DOK ratings of the extended benchmarks to those ratings given to the corresponding benchmarks in the Minnesota Academic Standards. Table 3.7 presents the percentage of extended benchmarks per grade-span 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 LAL criteria; however, it is reasonable to expect that as many as half of the extended benchmarks 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 benchmarks with performance expectations at a higher level than the regular content standards.

Table 3.7. Frequency of Extended Benchmarks at Same, Lower, or Higher Levels of Complexity Compared to Related Benchmarks

	Extended Benchmarks at Varying Levels of Complexity					
	Same		Lower		Higher	
Grade Level	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
5	3	50%	1	17%	2	33%
8	0	0%	3	50%	3	50%
High School	2	33%	4	67%	0	0%

For grade 5, panelists rated at least half of the extended benchmarks as assessing student knowledge at the same level of complexity as the benchmarks, while panelists for grade 8 rated half of the statements as lower than the benchmarks. The High School panelists rated all of the extended benchmarks as assessing the same or a lower level of complexity as the benchmarks. Panelists in grades 5 and 8 rated at least two extended benchmarks as higher in complexity than the related benchmarks. We identify those extended benchmarks rated with higher DOK ratings than the corresponding benchmarks in Table 3.8.

Table 3.8. Extended Benchmarks with Higher DOK Ratings than Corresponding Minnesota Academic Standards

Grade Level	Extended Benchmark	Identifier from MTAS Test Specifications
5	Students will identify and describe how states of matter change as a result of heating and cooling.	Grade 4, Strand 2 – Physical Science, Sub-strand 1 – Matter, 4.2.1.2.2
5	Students will identify methods of personal hygiene that help prevent germs from entering the body.	Grade 4, Strand 4 – Life Science, Sub-strand 4 – Human interactions with living systems, 4.4.4.2.1
8	Students will identify common engineered systems and evaluate their impact on the daily life of humans.	Grade 6, Strand 1 – The Nature of Science and Engineering, Sub-strand 2 – The practice of engineering, 6.1.2.1.1
8	Students will recognize the effects of balanced or unbalanced forces on an object.	Grade 6, Strand 2 – Physical Science, Sub-strand 2 – Motion, 6.2.2.2.1
8	Students will identify the effects of weathering, erosion, and deposition of sediment on landforms.	Grade 8, Strand 3 – Earth and Space Science, Sub-strand 1 – Earth structure and processes, 8.3.1.2.1

We also asked panelists to directly compare the written performance expectations in the extended benchmarks with the full content standards. Panelists evaluated the language of each extended benchmark 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 extended benchmark asks students to 'group' or 'categorize' based on traits, these expectations are parallel. If a benchmark expects students to 'identify and explain' while the extended benchmark asks students to 'identify' only, these expectations are partly similar. When students are asked to 'distinguish between' in the benchmark but the extended benchmark requires students to 'recognize', then the expectation for demonstrating knowledge is different. Table 3.9 shows the results of this comparison. At least 90% of the extended benchmarks should be rated as 'some' or 'all' compared with the full content standards.

Table 3.9. Mean Number of Extended Benchmarks at Various Levels of Performance Centrality

Grade	Performance Centrality Rating	Mean	SD	Percentage of Extended Benchmarks per Rating[a]
5	None	0.75	0.50	13%
	Some	2.25	1.26	38%
	All	2.25	0.50	38%
8	None	0.00	0.00	0%
	Some	0.50	1.00	8%
	All	5.25	1.50	88%
High School	None	0.75	0.96	13%
	Some	4.75	1.26	79%
	All	0.50	0.58	8%

[a] Total may sum to above 100% because percentages are based on mean numbers.

Only the extended benchmarks for grade 8 passed the minimum level of acceptability, with all of the content expectations rated as requiring the same or similar type of performance as the benchmarks. In grade 5 and High School, at least one extended benchmark received a rating of 'no similarity' to the corresponding benchmark.

> ***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. Since the Science MTAS assessments are grade-span tests, this criterion was not evaluated.

> ***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 extended benchmarks by providing two separate ratings. First, we asked panelists to determine the level of communication required by each extended benchmark in order for students to demonstrate knowledge. The common categories applied, according to the LAL method, include three ability levels for students with significant disabilities[7]:

- 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 benchmarks 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 Individualized Education Plans (IEPs). However, if the level of communication required in the extended benchmarks is always 'symbolic,' it becomes much more difficult for supports to be provided 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 benchmarks (and assessment tasks) be pre-symbolic. Thus, the minimum level of acceptability is that the access point for at least 90% of the extended benchmarks should be pre-symbolic.

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

Table 3.10 presents panelists' mean ratings on the communication levels needed to demonstrate content knowledge for each set of grade level extended benchmarks.

Table 3.10. Mean Number of Extended Benchmarks Rated at Each Level of Symbolic Communication

Grade	Level of Symbolic Communication Required	Mean	SD	Percentage of Extended Benchmarks per Rating [a]
5	Pre-symbolic	0.00	0.00	0%
	Early Symbolic	1.50	1.73	25%
	Full Symbolic	3.75	2.06	63%
8	Pre-symbolic	5.75	0.50	96%
	Early Symbolic	0.00	0.00	0%
	Full Symbolic	0.00	0.00	0%
High School	Pre-symbolic	1.25	0.50	21%
	Early Symbolic	2.00	0.82	33%
	Full Symbolic	2.75	0.50	46%

[a] Total may sum to above 100% because percentages are based on mean numbers.

Based on the panelists' ratings, only the grade 8 extended benchmarks met the minimum requirement of 90%. For High School, 21% of the extended benchmarks were rated as being pre-symbolic, and 0% of the extended benchmarks were rated as such in grade 5. These outcomes may indicate that students with the lowest level of symbolic abilities cannot access the full range of content expectations especially at grade 5 and High School. We encourage Minnesota to review the extended benchmarks to evaluate accessibility.

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 the extended benchmarks? Panelists provided a simple 'yes' (accessible to all) or 'no' (not accessible to some groups) response 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 3.11 includes the percentage of extended benchmarks that were judged as accessible to all groups.

Table 3.11. Frequency of Extended Benchmarks Rated as Accessible to All Students

Grade	Frequency	Percentage of Extended Benchmarks per Rating
5	5.25	88%
8	5.75	96%
High School	6.00	100%

In contrast to their ratings on symbolic communication abilities, panelists felt that the extended benchmarks across all grades could be accessed by a wide range of students with different physical and cognitive disabilities.

Summary and Discussion of Extended Benchmarks and Minnesota Academic Standards

For this alignment evaluation, panelists reviewed the extended benchmarks for Science in two ways. First, they evaluated the content alignment (Criteria 2 and 3 from the LAL method) between the grade span extended benchmarks 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 the majority of extended benchmarks across grade levels to link sufficiently with the Minnesota Academic Standards. In comparison, while most panelists found these content expectations to be accessible to various disability groups, they did express concern over access for students with limited symbolic communication.

Table 3.12 displays the overall conclusions regarding content alignment between the extended benchmarks and Minnesota Academic Standards for Science. These judgments are based on whether the extended benchmarks achieved acceptable levels of linkage with the full content standards for each set of grade level extended benchmarks.

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

Table 3.12. Summary Conclusions on Alignment of Extended Benchmarks to Minnesota Academic Standards for Science on LAL Criteria 2 and 3

Grade Level Tests	Criterion2	Criterion 3	
	Age Appropriate	Content Centrality	Performance Centrality
	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?
5	Partial	Partial	Partial
8	High	High	High
High School	Partial	Partial	Partial

The content alignment conclusions in Table 3.12 indicate that the grade-level Science extended benchmarks link well to the Minnesota Academic Standards in grade 8. However, there is only a partial linkage to the extended benchmarks in grade 5 and High School. In both grades, there was at least one extended benchmark that was rated as containing content that is off grade level. There were one to two extended benchmarks in grade 5 and High School that panelists rated as having a 'weak' or 'no' link to the Minnesota Academic Standards. Additionally, panelists rated at least one benchmark in grade 5 and High School as not meeting the same performance expectations as the Minnesota Academic Standards.

Table 3.13 displays the overall conclusions on content accessibility pertaining to Performance Accuracy (content accessibility) for the extended benchmarks. For this criterion, conclusions reflect overall judgments of acceptability based on access to the content expectations[8].

- 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 3.13. Summary Conclusions on Performance Accuracy (LAL Criterion 7) of Extended Benchmarks for Science

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

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

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

Chapter 4: Results: Science MTAS Tasks and Extended Standards

In this chapter, we report the results of panelists' ratings on the Science MTAS tasks per grade assessment. We present the results on the LAL Criteria 1 through 7. For grades 5 and 8, tasks were rated by four panelists in each group. There were four panelists in High School; however, data for one of the panelists was not captured for the task evaluation. All results for High School are based on three panelists.

Results on Science MTAS Tasks based on LAL Criteria

Ratings involved evaluation of the assessment relative to the extended benchmarks on all seven of the LAL criteria. Most results reflect mean ratings on a series of scales. Mean ratings were derived from frequency counts (per panelist) of how many extended benchmarks were rated at each response option. From these counts, we then calculated the mean number of extended benchmarks 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 LAL 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, mathematics, science).

Per the USDE (2005), alternate assessments counting towards Title I must assess students only on academic content, as opposed to functional life skills. The nature of the MTAS tasks has not changed since the previous alignment study conducted in 2008 evidenced that the majority of performance tasks were academic. For this reason, panelists did not rate this criterion providing more time to evaluate the other criteria. This criterion was generated by NAAC when it was not uncommon for alternate assessments to combine academic and functional life skills on the same test. More recent alternate assessments rarely include non-academic content, as they would represent construct irrelevant variance. Facilitators reviewed all MTAS tasks prior to the workshop and found that all MTAS tasks are academic.

> **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 4.1 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. The grade 8 and High School science tests surpassed the minimum requirement of 90% tasks rated as "adapted" or "neutral". This finding indicates that panelists found all of the tasks to be linked to grade-level content in those grades. In grade 5, the 90% minimum was just missed. There was one task that all panelists rated as measuring content off-grade level.

Table 4.1. Mean Percentage of Tasks at Various Levels of Age Appropriateness

Grade	Age Appropriateness Rating	Mean	SD	Percentage of Tasks per Rating
5	Adapted	7.75	0.50	86%
	Neutral	0.25	0.50	3%
	Inappropriate	1.00	0.00	11%
8	Adapted	8.75	0.50	97%
	Neutral	0.25	0.50	3%
	Inappropriate	0.00	0.00	0%
High School	Adapted	4.33	1.15	48%
	Neutral	4.33	0.58	48%
	Inappropriate	0.33	0.58	4%

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 extended benchmarks 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 extended benchmark. Table 4.2 shows that all raters considered all tasks for each grade science test to link to an extended benchmark.

Table 4.2. Mean Number of Tasks Linked to Extended Benchmarks

Grade	Percentage of Tasks Linked to Extended Benchmarks
5	100%
8	100%
High School	100%

We also asked panelists to evaluate *how well* the tasks targeted the extended benchmarks. At least 90% of tasks should be judged as moderately to closely linked with the extended benchmarks for acceptability. Table 4.3 presents the mean number and percent of tasks that fell into each category based on panelists' ratings.

Table 4.3. Mean Percent of Tasks at Various Levels of Content Centrality

Grade	Content Centrality Rating	Mean	SD	Percentage of Tasks per Rating
5	Not Link	0.25	0.50	3%
	Weak Link	1.50	1.00	17%
	Moderate Link	1.50	1.29	17%
	Close Link	5.75	1.50	64%
8	Not Link	0.00	0.00	0%
	Weak Link	0.25	0.50	3%
	Moderate Link	2.50	1.29	28%
	Close Link	6.25	0.96	69%
High School	Not Link	0.67	0.58	7%
	Weak Link	4.33	1.52	48%
	Moderate Link	3.67	1.53	41%
	Close Link	0.33	0.58	4%

Panelists in grade 8 rated the majority of tasks as linked to the target content of the extended benchmarks. There was at least one task in grade 5 that panelists rated as having a 'weak' or 'no' link to the extended benchmarks. A greater concern is in the High School ratings where more than half of the tasks were rated as having a 'weak' or 'no' link to the extended benchmarks.

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 extended benchmarks should require students to make some type of distinction. Table 4.4 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 4.4. Mean Percent of Tasks at Various Levels of Performance Centrality

Grade	Performance Centrality Rating	Mean	SD	Percentage of Tasks per Rating
5	All	3.75	1.26	42%
	Some	4.25	1.26	47%
	None	1.00	0.00	11%
8	All	7.25	0.96	81%
	Some	1.75	0.96	19%
	None	0.00	0.00	0%
High School	All	0.33	0.58	4%
	Some	6.67	1.53	74%
	None	1.00	1.00	11%

Grade 8 surpassed the minimum level of acceptability (90%) of tasks assessing students on at least some of the same performance expectations as the extended benchmarks. Grade 5 and High School had at least one task that panelists rated as not assessing students at the same performance expectations as the extended benchmarks resulting in less than the minimum level of acceptability.

> **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 extended benchmarks 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* extended benchmarks were assessed. Categorical concurrence describes the extent to which the extended benchmarks are covered by the assessment. For a general education 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 strand. This change is appropriate because alternate assessments tend to include considerably fewer items or tasks compared with a general education assessment. In addition, a single task may assess multiple content expectations[9].

Table 4.5 summarizes the MTAS alignment results for categorical concurrence. As Table 4.5 indicates, all grade-level tests met the Webb alignment criterion of at least one task per content strand.

Table 4.5. Summary of Categorical Concurrence Results for Science MTAS by Grade Level

	Mean Number of Tasks per Strand				
Grade Level	The Nature of Science & Engineering	Physical Science[a]	Earth & Space Science[a]	Life Science	Strands with at Least One Task
5	2.00	1.00	3.00	3.00	4 of 4
8	2.00	3.00	2.00	2.00	4 of 4
High School	1.00	n/a	n/a	8.00	2 of 2

[a] Not assessed in High School.

[9] The psychometric trade-off is that fewer tasks per strand may lead to a decrease in scoring accuracy.

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 extended benchmark using a rating scale (see Appendix C for the LAL DOK level descriptions). Next, as they reviewed performance tasks, panelists rated the level of processing needed to perform the task using the same DOK rating scales. Table 4.6 shows the mean percentage of tasks rated at each DOK level per grade level.

Table 4.6. Mean Percentage of Tasks at Each DOK Level

Grade	Task DOK Rating	Mean	SD	Percentage of Tasks per Rating
5	None	0.00	0.00	0%
	Attention	0.00	0.00	0%
	Memorize/recall	2.50	1.29	28%
	Performance	2.75	0.96	31%
	Comprehension	2.00	0.82	22%
	Application	1.50	0.71	8%
	Analysis, Synthesis, Evaluation	1.00	0.00	11%
8	None	0.00	0.00	0%
	Attention	0.00	0.00	0%
	Memorize/recall	4.25	0.50	47%
	Performance	2.67	1.53	22%
	Comprehension	2.00	1.00	17%
	Application	1.25	0.50	14%
	Analysis, Synthesis, Evaluation	0.00	0.00	0%
High School	None	0.00	0.00	0%
	Attention	0.00	0.00	0%
	Memorize/recall	3.67	2.08	41%
	Performance	1.00	0.00	11%
	Comprehension	2.33	1.53	26%
	Application	4.00[a]	n/a	15%
	Analysis, Synthesis, Evaluation	1.00	0.00	7%

[a] Only one reviewer gave this rating.

We then compared these two separate judgments about cognitive complexity (one for the extended benchmark, 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 4.7 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 (extended benchmarks), the table refers to consistency between the performance tasks and the extended benchmarks 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 strand. Webb's suggested criterion for this alignment indicator is the same as for a general education assessment – at least 50% of the tasks should have complexity

ratings at or above the level of the corresponding extended benchmark per content strand. The percentages that do not reach the 50% criteria are bolded, and the strands are noted in the far right column.

Table 4.7. Summary of Depth-of-Knowledge Results for Science MTAS by Grade Level

	Percent of Tasks with DOK At or Above the Level of the Extended Benchmarks per Strand					
Grade Level	The Nature of Science & Engineering	Physical Science[a]	Earth & Space Science[a]	Life Science	Number of Strands Assessed Adequately	Specific Strands Assessed Inadequately
5	100%	**25%**	**33%**	92%	2 of 4	Physical Science; Earth & Space Science
8	**0%**	100%	**38%**	50%	2 of 4	The Nature of Science & Engineering; Earth & Space Science
High School	**0%**	n/a	n/a	67%	1 of 2	The Nature of Science & Engineering

[a] Not assessed in High School.

All of the grade-level Science assessments met Webb's target of 50% consistency with the extended benchmarks for the Life Science strand. Additionally, Webb's target was met for grade 5 Nature of Science and Engineering strand as well as grade 8 Physical Science. For these grade and strand combinations, a number of tasks assessing students at the same cognitive complexity level as the extended benchmarks for each strand existed. However, for other grade and strand combinations this did not occur. In grade 5, panelists' found tasks assessing extended benchmarks under the strands Physical Science and Earth and Space Science to be lower in complexity than expected. In grade 8, panelists' ratings on the Nature of Science and Engineering and Earth and Space Science content strands are lower in cognitive complexity than expected. Finally, panelists' ratings in High School showed no tasks at the appropriate cognitive complexity level in Nature of Science and Engineering. Based on these findings, Minnesota may wish to review the performance tasks and extended benchmarks to determine if changes are necessary for one or both.

Range of Knowledge Correspondence. Range of knowledge correspondence measures how fully the tasks cover each of the extended benchmarks within each strand. The assessed extended benchmarks 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 extended benchmarks must be matched to one or more tasks.

Table 4.8 summarizes the range-of-knowledge correspondence results for each grade level of the MTAS. We computed the number of extended benchmarks covered for each strand separately for each panelist and then averaged across panelists to obtain the summary alignment indicator.

Table 4.8. Summary of Range-of-Knowledge Results for Science MTAS by Grade Level

Grade Level	Percent of Extended Benchmarks per Strand Matched to at Least One Task				Number of Strands Assessed Adequately	Specific Strands Assessed Inadequately
	The Nature of Science & Engineering	Physical Science[a]	Earth & Space Science[a]	Life Science		
5	100%	100%	100%	100%	4 of 4	
8	100%	100%	100%	100%	4 of 4	
High School	100%	n/a	n/a	100%	2 of 2	

[a] Not assessed in High School.

As Table 4.8 demonstrates, all of the strands were covered adequately by the assessments across all grade levels. That is, at least 50% of the extended benchmarks were linked to at least one performance task.

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

The content balance is determined by calculating an index, or score, for each strand[10]. According to Webb, the minimum acceptable index for a single strand is 70 (on a scale of 0 to 100, with 100 representing perfect balance). To be clear, a strand 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 4.9 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.

Table 4.9. 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
	The Nature of Science & Engineering	Physical Science[a]	Earth & Space Science[a]	Life Science		
5	100	100	83	83	4 of 4	
8	100	83	100	100	4 of 4	
High School	100	n/a	n/a	83	2 of 2	

[a] Not assessed in High School.

[10] 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.

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. Since the Science MTAS assessments are grade-span tests, this criterion was not evaluated.

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

The sixth LAL criterion pertains to demonstration of student learning. 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 4.10 includes the group consensus ratings on the degree of student inference evident in the Science MTAS assessment per grade level.

Table 4.10. Degree of Inference Evident on Student Learning in Science MTAS Assessments

Grade	Dimensions	Rating	Rationale
5	Level of Accuracy		There is no inference for students who are getting all or mostly 1's. If a student received a variety of 2's and 3's, then there can be an inference made.
	Level of Independence	H	When we administer this test, the student is shown the response, additional support is given only if the answer is wrong.
	New Learning	L	It is one given test without a pre-test or baseline. This test is not a measure of growth.
	Generalization across People and Settings		Does not match the test.
	Generalization across Materials and Activities	H	They used a variety of settings and materials, but so many different pictorial presentations of the same item could be confusing. They show only white boys doing the tasks (with exception of a woman shopping), and there are no visually handicapped people (wheelchairs) in the presentation pages.
	Standard Setting	L	A student could show proficiency if they guessed and got a few right while the rest of the tasks needed prompting after getting it wrong or for a wrong answer.
	Program Quality Indicators	H	No community support or school programming that influences the student achievement on the test.
8	Level of Accuracy	H	The scoring procedure will show evidence of student learning.
	Level of Independence	H	The students are able to independently respond through a variety of communication modes.
	New Learning	H	We don't have a baseline or pre-test. It is a one-time assessment that includes grade level differentiation.
	Generalization across People and Settings	H	There are different cultures and genders represented. There are a number of settings presented as well.
	Generalization across Materials and Activities	L	All standards have more than one task. Some items approach getting at content in the same way.
	Standard Setting	H	The standard setting procedure is based on independent student performance and a high level of accuracy
	Program Quality Indicators	H	There are no other indicators factored into the student score.

Table 4.10. Degree of Inference Evident on Student Learning in Science MTAS Assessments (continued)

Grade	Dimensions	Rating	Rationale
	Level of Accuracy	N	Students get points for wrong answers. If these points count toward proficiency, then there has been no student inference.
	Level of Independence	L	A score of 2 shows added guidance to aid the student.
	New Learning	L	There is no baseline and the progression of learning from grade level to grade level is unclear.
High School	Generalization across People and Settings	L	Some of the pictures are misleading.
	Generalization across Materials and Activities	L	One benchmark had only one task and others had multiple tasks, but the tasks were quite similar to each other.
	Standard Setting		We are unclear of the proficiency level.
	Program Quality Indicators	H	There are no program indicators involved.

H = high student inference; L = low student inference; N = no student inference

For grade 8, panelists determined that the scoring rubric and achievement standards allowed for clear distinctions among student responses on tasks, as demonstrated by their ratings on Level of Accuracy. In High School, panelists determined that there was no distinction among student responses, and in grade 5, panelists were torn in providing a rating as their rationale points to a conditional response. Panelists' ratings suggest that the grade 5 assessment seems to enable students to demonstrate learning better overall (across multiple dimensions) compared to the higher grades (8 and High School). Panelists noted that it is difficult to identify evidence of "new learning" because no baseline exists for comparison, particularly if tasks between grades are quite similar.

> **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. 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 accommodations or supports can be provided for different types of students without substantially altering the target content.

Table 4.11 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. Grade 8 met this minimum requirement. While the panelists for grade 5 and High School rated none to half of tasks as requiring pre-symbolic ability, they did rate between 4 to 6 tasks as requiring students to possess full symbolic or early symbolic ability to comprehend and respond.

Table 4.11. Mean Percentage of Tasks at Various Levels of Symbolic Communication

Grade	Symbolic Ability Rating	Mean	SD	Percentage of Tasks per Rating
5	Pre-symbolic	0.00	0.00	0%
	Early Symbolic	6.00	1.15	67%
	Symbolic	3.00	1.15	33%
8	Pre-symbolic	9.00	0.00	100%
	Early Symbolic	0.00	0.00	0%
	Symbolic	0.00	0.00	0%
High School	Pre-symbolic	4.33	3.79	48%
	Early Symbolic	4.67	3.79	52%
	Symbolic	0.00	0.00	0%

Concerning the accessibility of task content to students with a variety of disabilities, panelists for the grade 8 and High School assessments considered the majority of operational tasks to be accessible to a wide range of students, as shown in Table 4.12. In grade 5, panelists found that only about a quarter of tasks were accessible to a wide range of students.

Table 4.12. Frequency of Tasks Rated as Accessible to All Students

Grade	Frequency	Percentage of Tasks per Rating
5	2.50	28%
8	9.00	100%
High School	8.33	93%

Finally, panelists evaluated the Science MTAS tasks on an additional dimension under Criterion 7. A common approach to administering an alternate assessment is for teachers to offer accommodations or supports (i.e., assistive technology; prompts if needed) as appropriate for a given student. Panelists rated each task as to whether they could in fact be accommodated or supports offered, particularly without altering the target of the assessment. Table 4.13 includes the mean number of tasks panelists found amenable to these types of changes.

Table 4.13. Frequency of Tasks Rated as Amenable to Accommodations or Supports

Grade	Frequency	Percentage of Tasks per Rating
5	6.25	69%
8	9.00	100%
High School	5.67	63%

At grade 8, panelists found that all tasks could be altered appropriately for individual students. In grade 5 and High School, panelists found that about six of the nine tasks could be altered appropriately.

Reliability Results

In this section, we report on agreement analyses on panelists' ratings. Primarily, we compare panelists' ratings on content match to the test contractor's intended content match.

Panelist-Test Developer Analyses

We assessed the agreement between our panelists' judgments of assessed content and the AIR item specifications for each performance task. Such a comparison provides an independent evaluation of the test contractor's content assignment. Table 4.14 includes these comparisons by noting the percent of tasks with exact agreement between panelists and AIR on content match for operational tasks. Panelists agreed completely for the grade 8 tasks. Panelists showed moderate agreement with the test contractor for the grade 5 and High School assessments. For one of the grade 5 tasks, the extended benchmark assessed is at the fifth grade level while panelists noted it was the third grade level that was assessed instead. For a second task in grade 5, half of the reviewers chose a different extended benchmark than the one specified. One task on the High School assessment was listed as assessing the Nature of Science and Engineering strand; however, all panelists reported an extended benchmark assessing the Life Science strand instead. A second task on the High School assessment had a mismatch between the intended extended benchmark assessed and the one assigned by panelists.

Table 4.14. Percentage Agreement between Panelists and AIR on Assessment Target for MTAS Operational Science Tasks

Grade	Number of Operational Tasks	Number of Raters	Percentage of Tasks with Exact Agreement
5	9	4	83%
8	9	4	100%
High School	9	3	81%

Summary and Discussion of Science MTAS Tasks and Extended Standards

Table 4.15 displays the overall conclusions regarding content alignment between the Science MTAS assessments and the extended standards. These judgments are based on whether the extended standards achieved acceptable levels of linkage with the full content standards for each grade test. The minimum level for each of the criteria in Table 4.15 is 90%.

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

Table 4.15. Summary Conclusions on Alignment of Science MTAS Assessments to Extended standards for LAL Criteria 2, 3, and 4

Grade Level Tests	Criterion2	Criterion 3		Criterion 4	
	Age Appropriate	Content Centrality	Performance Centrality	Content Coverage	
	Is task content referenced to student's assigned grade level?	Do tasks link to the target content in the Extended standards?	Does the performance of task link to expectations of the Extended standards?	Do the tasks assess students at the appropriate breadth of knowledge? [a]	Do the tasks assess students at the appropriate depth-of-knowledge? [b]
5	Partial	Partial	Partial	High	Partial
8	High	High	High	High	Partial
High School	High	Weak	Partial	High	Partial

[a] Conclusions are 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.
[b] Conclusions are based on the results from the DOK consistency analyses.

As Table 4.15 illustrates, the 2012 Science MTAS assessments linked well to the content of the extended standards on the majority of dimensions, particularly for grade 8. In grade 5, the minimum level for 'High linkage' was just missed on age appropriateness and performance centrality. For these two criteria, panelists reported as least one task that was not age appropriate or linked to the expectations of the extended benchmarks. There is cause for concern in High School as the link between tasks and extended benchmarks was clearly seen by panelists for only half of the tasks. On depth-of-knowledge assessed, panelists determined that many tasks assessed students at a different level of cognitive complexity than expected in the extended benchmark for all grades.

Table 4.16 includes results relative to Criterion 6 and 7 of the LAL method. These rating questions asked panelists to determine whether the assessment tasks are designed in such a way that students can demonstrate knowledge at various levels of functioning and ability. Ratings in this case are based on evaluations of accessibility, rather than on content alignment.

- Excellent - all tasks are acceptable
- Good - most tasks are acceptable (at least 90%)
- Acceptable - many tasks are acceptable (70%-89%)
- Questionable - few tasks are acceptable (less than 70%)

Table 4.16. Summary Conclusions on Accessibility (LAL Criteria 6 and 7) of Science MTAS Assessments

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

The most noticeable issue regarding accessibility for the Science MTAS assessments concerns the measurement of achievement (Criterion 6), particularly for grade 5 and High School. For these grades, panelists indicated difficulty in being able to make inferences about student achievement regarding level of accuracy, generalizability, and standard setting. Panelists for all grades raised concerns over a need for clear guidelines explaining the use of appropriate additional supports, as well as an inability to truly infer evidence of new learning with the MTAS due to lack of a baseline measure.

Results on Criterion 7 were mixed. For grade 8, excellent results can be found for all questions related to performance accuracy. However, grade 5 and High School showed questionable results. It is worthwhile to review tasks in these two grades on the level of symbolic ability required of students to ensure that all students can access the content and demonstrate what they know.

Chapter 5: 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 benchmarks and benchmarks. 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.

Two types of alignment evaluations were applied to the Grades 5, 8, and High School Science MTAS tests: (a) alignment of the extended benchmarks to the Minnesota Academic Standards for Science and (b) alignment of the Science MTAS assessments to the extended benchmarks. The cumulative results point to reasonable content linkage of the assessment and extended benchmarks with the content standards on most dimensions. The findings on content accessibility were mixed for the assessments. For the 2012 assessments, panelists identified some concerns over the ability to make accurate inferences about student knowledge based on the present scoring rubric and achievement standards. Finally, panelists noted that for some content expectations may be set too high which could exclude students with pre-symbolic or even early symbolic communication ability to access the assessment.

HumRRO does suggest areas where Minnesota could improve access to the content of the extended benchmarks, as well as strengthen content alignment and access for the Science MTAS assessments. 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 3.12 and 3.13 in Chapter 3 and Tables 4.15 and 4.16 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 extended benchmarks or assessment fall short on a serious issue, such as accessibility.

Extended Benchmarks for Science

- ***Review the access points for the extended benchmarks and performance tasks at grade 5 and High School.*** For these grades, panelists identified some extended benchmarks and performance tasks that may limit access only to those students with higher symbolic abilities, thus excluding a portion of students from the assessment system. Reviewing the extended benchmarks and performance tasks may involve additional bias reviews to modify the current expectations; or, additional explanation (e.g., content limitations, examples) within the MTAS Test Specifications document may be sufficient to better illustrate how teachers might make these content expectations more appropriate for students with lower symbolic abilities.
- ***Review the link between the extended benchmarks/performance tasks and the Minnesota Academic Standards in grade 5 and High School.*** Panelists identified one to two extended benchmarks as having a ‘weak’ or ‘no’ link to the Minnesota Academic Standards. In both grades, there were two panelists who rated one extended benchmark as having ‘no’ link while the other two panelists rated it as ‘weak’. One issue highlighted by High School panelists was that the extended benchmarks differed in the expectations for

students. For example, the benchmark may ask students how something will change while the extended benchmark asks students to recognize the change. We suggest that MDE review the extended benchmarks and performance tasks to ensure that the content expectations are similar to those in the Minnesota Academic Standards.

MTAS Performance Tasks

- ***Review performance tasks for depth-of-knowledge required of students relative to the extended benchmarks.*** Panelists determined that at least half of the performance tasks did not assess student knowledge related to several Science content strands at levels of cognitive complexity comparable to those expected in the extended benchmarks. We suggest that MDE review *both* the extended standards and the tasks with content and special education experts to increase the agreement between the depth-of-knowledge expected of students and that which is assessed. Assessments should always reflect the content expectations, but it may be worthwhile to confirm that the extended benchmarks are written at the appropriate level of cognitive demand for Minnesota students.
- ***Improve the ability of the Science MTAS assessments to accurately demonstrate student knowledge.*** Panelists rated the assessment materials as providing little to no inference about student learning on multiple dimensions. Furthermore, two general comments showed up in all grade level panels. First, panelists expressed concerns that extended benchmarks often had no more than one associated performance task, and generally, when two were used, the performance tasks were similar to each other. This is somewhat understandable given that there are six extended benchmarks and nine performance tasks on each assessment. In addition, panelists noted that it is difficult to identify evidence of "new learning" because no baseline exists for comparison, particularly if tasks between grades are quite similar.

  The issue of support is one that many states struggle with in trying to balance the use of a standardized assessment against allowing considerable flexibility for individual students with a broad range of disabilities. One option used in some states (e.g., Kentucky) is to include a checklist of common and appropriate accommodations and supports along with the conditions under which each accommodation can be provided. The Accommodations Manual, developed by the ASES SCASS, is a good resource that includes condition charts that can be adapted by each state (Thompson, Morse, Sharpe, & Hall, 2005). Although most test administrators are familiar with appropriate accommodations, research by NAAC suggests that competencies required to administer Alternate Assessments can vary widely, even within a state.

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

Minnesota Department of Education (January, 2008). *Minnesota test of academic skills (MTAS): Test specifications for reading and 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

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.

Thompson, S. J., Morse, A. B., Sharpe, M., & Hall, S. (2005). *Accommodations manual: How to select, administer, and evaluate use of accommodations for instruction and assessment of students with disabilities.* CCSSO State Collaborative on Assessment and Student Standards Assessing Special Education Students. Washington, DC. Retrieved from http://www.osepideasthatwork.org/toolkit/accommodations_manual.asp

U.S. Department of Education. (August, 2005). *Alternate achievement standards for students with the most significant cognitive disabilities.* Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education. Retrieved from http://www.ed.gov/admins/lead/account/saa.html#guidance.

Webb, N. L. (2005). *Webb alignment tool: Training manual.* Madison, WI: Wisconsin Center for Education Research. Available: http://www.wcer.wisc.edu/WAT/index.aspx.

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in Science and Math education.* Washington, D.C.: Council of Chief State Schools Officers.

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

Appendix A.
Webb Alignment Results per Grade Level Assessment

Webb Alignment Results

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

Categorical Concurrence

The categorical concurrence results for grades 5, 8, and High School of the Science MTAS assessment are presented below. Each table includes: the target number of tasks from the test blueprint; the mean number of tasks matched by panelists; the standard deviation among panelists' ratings; and, the final alignment conclusion (Yes or No). The bottom row indicates the percentage of strands that met the minimum alignment criterion.

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

	Number of Tasks per Strand			
Title of Strand	Target # Tasks from Blueprint	Mean Tasks Matched	Standard Deviation	At Least One Task per Strand
The Nature of Science & Engineering	1-2	2.00	0.00	Yes
Physical Science	1-2	1.00	0.00	Yes
Earth & Space Science	2-4	3.00	0.00	Yes
Life Science	2-4	3.00	0.00	Yes
Total	9			
Percentage 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

	Number of Tasks per Strand			
Title of Strand	Target # Tasks from Blueprint	Mean Tasks Matched	Standard Deviation	At Least One Task per Strand
The Nature of Science & Engineering	1-2	2.00	0.00	Yes
Physical Science	2-4	3.00	0.00	Yes
Earth & Space Science	1-2	2.00	0.00	Yes
Life Science	2-4	2.00	0.00	Yes
Total	9			
Percentage 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

	Number of Tasks per Strand			
Title of Strand	Target # Tasks from Blueprint	Mean Tasks Matched	Standard Deviation	At Least One Task per Strand
The Nature of Science & Engineering	1-2	1.00	0.00	Yes
Life Science	5-8	8.00	0.00	Yes
Total	9			
Percentage of strands with at least one task: 100%				

Depth-of-Knowledge Consistency

The Depth-of-Knowledge (DOK) consistency results for grades 5, 8, 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 tasks. The tables include the mean percentage of tasks 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 tasks 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

		Depth-of-Knowledge Consistency						
Title of Strand	Mean Tasks per Strand	% Tasks Below		% Tasks Same Level		% Tasks Above		DOK Consistency Target Met
		M	S.D.	M	S.D.	M	S.D.	
The Nature of Science & Engineering	2.0	0.0	0.0	87.5	25.0	12.5	25.0	Yes
Physical Science	1.0	75.0	50.0	25.0	50.0	0.0	0.0	No
Earth & Space Science	3.0	66.7	0.0	33.3	0.0	0.0	0.0	No
Life Science	3.0	8.3	16.7	50.0	19.2	41.7	16.7	Yes
Percent of strands with 50% of task DOK at or above objective DOK: 50%								

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.	
The Nature of Science & Engineering	2.0	100.0	0.0	0.0	0.0	0.0	0.0	No
Physical Science	3.0	0.0	0.0	33.3	27.2	66.7	27.2	Yes
Earth & Space Science	2.0	62.5	25.0	37.5	25.0	0.0	0.0	No
Life Science	2.0	50.0	0.0	12.5	25.0	37.5	25.0	Yes
Percent of strands with 50% of task DOK at or above objective DOK: 50%								

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.	
The Nature of Science & Engineering	1.0	100.0	0.0	0.0	0.0	0.0	0.0	No
Life Science	8.0	33.3	14.4	25.0	12.5	41.7	26.0	Yes
Percent of strands with 50% of task DOK at or above objective DOK: 50%								

Range-of-Knowledge Correspondence

The results for Range-of-Knowledge correspondence for grades 5, 8, and High School of the Science MTAS assessment are presented below. The tables include the mean number, standard deviation, and percentage of extended benchmarks by strand. For acceptable range-of-knowledge correspondence, a minimum of 50% of content Extended standards within each strand should be matched to at least one task.

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

Title of Strand	Number of Extended Benchmarks	Mean Tasks per Strand	Range of Extended Benchmarks: Extended Benchmarks with At Least One Task M	Range of Extended Benchmarks: Extended Benchmarks with At Least One Task S.D.	Range of Extended Benchmarks: % of Total Extended Benchmarks per Strand M	Range-of-Knowledge Target Met
The Nature of Science & Engineering	1	2.0	1.0	0.0	100%	Yes
Physical Science	1	1.0	1.0	0.0	100%	Yes
Earth & Space Science	2	3.0	2.0	0.0	100%	Yes
Life Science	2	3.0	2.0	0.0	100%	Yes
Total	6					
Percentage of strands with 50% of extended standards linked to at least one task: 100%						

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

Title of Strand	Number of Extended standards	Mean Tasks per Strand	Range of Extended standards: Extended standards with At Least One Task M	Range of Extended standards: Extended standards with At Least One Task S.D.	Range of Extended standards: % of Total Extended standards per Strand M	Range-of-Knowledge Target Met
The Nature of Science & Engineering	1	2.0	1.0	0.0	100%	Yes
Physical Science	2	3.0	2.0	0.0	100%	Yes
Earth & Space Science	1	2.0	1.0	0.0	100%	Yes
Life Science	2	2.0	2.0	0.0	100%	Yes
Total	6					
Percentage of strands with 50% of extended standards linked to at least one task: 100%						

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

Title of Strand	Number of Extended standards	Mean Tasks per Strand	Range of Extended standards: Extended standards with At Least One Task		Range of Extended standards: % of Total Extended standards per Strand	Range-of-Knowledge Target Met
			M	S.D.	M	
The Nature of Science & Engineering	1	1.0	1.0	0.0	100%	Yes
Life Science	5	8.0	5.0	0.0	100%	Yes
Total	6					
Percentage of strands with 50% of extended standards linked to at least one task: 100%						

Balance-of-Knowledge Representation

The results for Balance-of-Knowledge representation for grades 5, 8, and High School of the Science MTAS assessment are presented below. The tables also include the percentage of tasks linked to each strand. The minimum acceptable balance index is a 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: Extended Benchmarks per Strand	Mean Extended Benchmarks 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.	
The Nature of Science & Engineering	1	1.0	2.0	22%	100	0.0	Yes
Physical Science	1	1.0	1.0	11%	100	0.0	Yes
Earth & Space Science	2	2.0	3.0	33%	83	0.0	Yes
Life Science	2	2.0	3.0	33%	83	0.0	Yes
Total	6						
Percentage of strands 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						
	Extended Benchmarks per Strand	Mean Extended Benchmarks 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.	
The Nature of Science & Engineering	1	1.0	2.0	22%	100	0.0	Yes
Physical Science	2	2.0	3.0	33%	83	0.0	Yes
Earth & Space Science	1	1.0	2.0	22%	100	0.0	Yes
Life Science	2	2.0	2.0	22	100	0.0	Yes
Total	6						
Percentage of strands 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						
	Extended Benchmarks per Strand	Mean Extended Benchmarks 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.	
The Nature of Science & Engineering	1	1.0	1.0	11%	100	0.0	Yes
Life Science	5	5.0	8.0	89%	83	4.3	Yes
Total	6						
Percentage of strands with a balance of representation index of 70 or greater: 100%							

Appendix B.
Summary of Panelist Comments on Tasks

Tables B.1 through B.3 present a synopsis of panelists' comments on the individual tasks of the Science MTAS. To maintain test security, individual task identifiers are not presented, nor are any comments that would reveal the content of a task. Column 2 indicates the number of tasks receiving such comments, and Column 3 reports how many panelists included this type of comment.

Table B.1. Grade 5 Science MTAS: Summary of Panelists' (N=4) Comments on Tasks by Topic

Comment	Number of tasks with comment	Number of panelists with comment
• Presentation pages and/or response cards are misleading, ambiguous, or difficult to make out.	2	3
• Task cannot be effectively modified for blind students.	2	3
• Task content is inappropriate for low cognitive students.	2	2
• Task cannot be modified.	2	2
• Task content does not align with the benchmark.	1	3
• Task is a poor example of knowledge demonstration.	1	1

Table B.2. Grade 8 Science MTAS: Summary of Panelists' (N=4) Comments on Tasks by Topic

Comment	Number of tasks with comment	Number of panelists with comment
• Task content contains inaccurate information.	2	3
• Task asks students to "evaluate the impact," which brings down the DOK.	2	1
• Presentation pages and/or response cards are misleading, ambiguous, or difficult to make out.	1	4
• Wording of the task or response options is confusing or misleading.	1	3
• Task content does not align with the benchmark.	1	1
• Task is similar to another task.	1	1

Table B.3. High School Science MTAS: Summary of Panelists' (N=4) Comments on Tasks by Topic

Comment	Number of tasks with comment	Number of panelists with comment
• Pictures on the presentation page are inaccurate, confusing, or misleading.	2	2
• Task wording is confusing.	1	2
• Task content does not align with the benchmark.	1	1
• More than one correct response	1	1
• Rigor of question changes when script changes from a 3-point task to a 2-point task.	1	1
• Task can be answered without using the corresponding graph.	1	1
• Pictures on the presentation page do not correspond to task.	1	1
• Response options contain a poor distractor.	1	1

Appendix C.
Sample Alignment Review Materials

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

MTAS Science
Panelist Instructions

	Rating Step	Documents Needed	File Format
1	DOK of MTAS extended benchmark (Consensus)	(1) MTAS Extended Benchmarks (HumRRO coded) (2) MTAS Science Ext Bench Consensus Response Sheet (3) "MTAS Science Ext Bench DOK_GradeX"	Print Copy Print Copy Excel file
2	DOK for MCA standards (Consensus)	(1) MCA Science Standards (HumRRO Coded) (2) MCA DOK Rating Response Sheet (3) "MCA Science Standards Consensus Rating_Gr X"	Print copy Print Copy Excel file
3	MTAS extended benchmark evaluation (Individual)	(1) MTAS Extended Benchmark (HumRRO coded) (2) MCA Science Standards (HumRRO Coded) (3) "MTAS Science Ext Bench_GradeX"	Print Copy Print Copy Excel file
4	MTAS tasks (Individual)	(1) MTAS Extended Benchmark (HumRRO coded) (2) MTAS item documents (e.g., presentation pages, response cards) (3) MTAS task administration manual (4) "MTAS Science Task Rate_GradeX"	Print copy Print copy Print copy Excel file
5	Student learning evaluation across grades (Individual)	(1) MTAS Extended Benchmark (HumRRO coded) (2) MTAS item documents (e.g., presentation pages, response cards) (3) MTAS task administration manual (4) "MTAS Science Student Learn_GradeX"	Print copy Print copy Print copy Excel file
6	Whole test evaluation across grade span (Individual, discuss as group-summarize)	(1) MTAS extended benchmark (HumRRO Coded) (2) "MTAS Science Whole Test_GradeX"	Print copy Excel file

1 Rate DOK for MTAS Extended Benchmark (Consensus)

Using the MTAS Science Extended Benchmark printout, assign a depth-of-knowledge (DOK) rating to each extended benchmark on the rating form provided. You will first do this independently and then your panel will come to consensus on each rating (3/4 majority, if necessary). The consensus ratings will be retained for analysis and input into the "MTAS Science Ext Bench DOK_Grade**X**" excel spreadsheet by the group leader. Change the file name to include group leader's 3 initials (example: MTAS Science Ext Bench DOK_Grade**X** _rcd).

DOK	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)

2 Rate DOK for MCA standards (Consensus)

Using the "MCA Science Standards'" printout, assign a depth-of-knowledge rating to each standard benchmark on the provided rating form. You will first rate the standards independently and then your panel will come to consensus on each rating (3/4 majority, if necessary). The consensus ratings will be retained for analysis and input into the "MCA Science Standards Consensus Rating_Gr**X**'" excel spreadsheet by the group leader. Change the file name to include group leader's 3 initials (example: MCA Science Standards Consensus Rating_Gr**X**'_rcd).

3 Rate the MTAS Extended Benchmarks (Individual)

Open the Excel file "MTAS Science Ext Bench _Grade**X**". Evaluate the extended benchmarks on all of the dimensions (columns) in the form. The extended benchmarks are rated on each dimension independently. We will discuss discrepant ratings on other dimensions if there is time.

A. Determine which MCA Standard best matches with the listed MTAS extended benchmarks. Enter the HumRRO standard ID found on the MCA standards document. If there is absolutely no MCA standard you can list, leave the cell blank.

B. Indicate *how well* you think that the MTAS extended benchmark actually links to the MCA standard (Content Centrality). Please reserve the use of a code of '1' (No Link) for two circumstances: (1) the extended benchmark does not link to any standard/benchmark, (2) the extended benchmark does not link to the standard listed but is within the overall standard category.

Category	Code	Description
Content Centrality	1	No link
	2	Weak link
	3	Moderate link
	4	Close link

C. Determine to what extent the extended benchmark measures student performance expected in the MCA standard.

Category	Code	Description
Performance Centrality	N	None - performance expectation is different from content standard
	S	Some - performance expectation partially matches content standard (content standard may include two different performance expectations, such as *'identify and explain'*).
	A	All - performance expectation is identical to content standard

D. Evaluate whether the extended benchmark is appropriate for the grade-level at which the content is measured. Content may be grade-level appropriate, off grade-level, or grade-level neutral (meaning that the content/topic could be assessed at any grade).

Category	Code	Description
Grade-Level Appropriate	I	Inappropriate; off grade-level content
	N	Neutral; content is not grade-level bound
	A	Adapted from grade-level content

E. 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 <u>lowest</u> functioning student who could access this task.**

Category	Code	Description
Barriers to Demonstrating Knowledge		
Symbolic Communication	A	Awareness/Pre-symbolic (gesture, purposeful moving toward object/sound)
	E	Early Symbolic
	S	Symbolic (pictures, symbols, signs, speech)

F. Evaluate the accessibility of the extended benchmark for various disability groups. If the content 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.

Category	Code	Description
Barriers to Demonstrating Knowledge		
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).

4 Rate MTAS Tasks

Open the Excel file "MTAS Science Task Rate_Grade**X**" and add "_your 3 initials". You will be rating each individual task independently. We will discuss discrepant ratings. The ratings are similar to what you used for the extended benchmarks in Rating Steps 2 and 3 with two exceptions: a) having a Secondary extended benchmark match (rarely needed), and b) a third category of modification under Barriers (Below). Refer to the code descriptions in Sections 1 and 3 above.

Category	Code	Description
Barriers to Demonstrating Knowledge		
Modification	Y	Yes, the task can be modified without changing difficulty or meaning.
	N	No, modification will change the difficulty or meaning. (PLEASE provide annotation in Notes to explain).

5 Rate student learning (Performance Level Descriptors and Task Specifications)

Open the "MTAS Science Student Learn_Grade**X**", add "_your 3 initials". Rate the Task dimensions after reviewing item documents, scoring guidance, and the task administration manual, with regard to 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.

Degree of Inference about Student Learning
(based on scoring for each Alternate Assessment (AA) item/task or found in standard setting documentation)

Criterion	High Student Inference Can clearly infer student showed learning	Low Student Inference Student performance mixed with educator or program performance	No Student Inference Can clearly infer student did not have to show any learning, it is educator or program performance rated
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 or 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 or 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 or tasks by grade level (criteria 5).	One time performance AND grade level differentiation of AA items or tasks was not clear (criteria 5).	No baseline, pretest, and weak differentiation across grade level AA items or tasks suggest student could achieve proficiency by making same response year after year (criteria 5).
Generalizations across people and settings (Note: this is less important than conceptual generalization)	Items or tasks are demonstrated across people or settings for full credit.	At least some items or tasks are demonstrated across more than one person or setting.	Item or task is only demonstrated with one person in one setting.
Generalizations across materials and activities (conceptual generalization)	Items or tasks are demonstrated across materials and activities or all standards have more than one item or task.	At least some items or tasks are demonstrated across materials or activities; or there is more than one item or task for some standards.	Item or task is only demonstrated with one specific materials and activity, there is only one item or 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.

6 Rate ‘Whole Test’ barriers to demonstrating student knowledge

Open the Excel “MTAS Science Whole Test_Grade**X**” file and add “_your 3 initials”. 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. If N (no) response is given, provide an explanation for the rating.

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 for grade 5 as an example.

Grade	Strand	Substrand	Standard	Code	Benchmark	HumRRO ID Number
5	1. The Nature of Science and Engineering	1. The Practice of Science	1. Science is a way of knowing about the natural world, is done by individuals and groups, and is characterized by empirical criteria, logical argument and skeptical review.	5.1.1.1.1	Explain why evidence, clear communication, accurate record keeping, replication by others, and openness to scrutiny are essential parts of doing science.	51111
5				5.1.1.1.2	Recognize that when scientific investigations are replicated they generally produce the same results, and when results differ significantly, it is important to investigate what may have caused such differences. For example: Measurement errors, equipment failures, or uncontrolled variables.	51112
5				5.1.1.1.3	Understand that different explanations for the same observations usually lead to making more observations and trying to resolve the differences.	51113
5				5.1.1.1.4	Understand that different models can be used to represent natural phenomena and these models have limitations about what they can explain. For example: Different kinds of maps of a region provide different information about the land surface.	51114

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 Science Tasks: Grade 3-5

MTAS Task	DOK	Match Task to Extended Benchmark	Content Centrality	Performance Centrality	Task Match to Secondary Extended Benchmark	Grade-Level Appropriate	Barriers to Demonstrating Knowledge			Notes/Comments
	What is the DOK level of this Task? 0-None 1-Attention 2-Recall 3-Performance 4-Comprehension 5-Application	The listed Task best matches with which MTAS Extended Benchmark? Enter the HumRRO ID number of the MTAS Extended Benchmark (5-digit number)	How well does the Task link to the MTAS Extended Benchmark? 1- No link 2- Weak link 3- Moderate link 4- Close link	Does the Task measure performance of MTAS Extended Benchmark? N - None S - Some A - All	Which secondary MTAS Extended Benchmark does the Task assess? Enter HumRRO ID number Note: Should rarely be needed; provide an explanation if used	Is the Task grade-level appropriate? I - Inappropriate N - Neutral A - Adapted	What level of symbolic communication does the Task require? Y - Yes N - No	Is the Task accessible to different disability groups? Y - Yes N - No	Can the Task be modified/supports provided without changing meaning or difficulty? Y - Yes N - No	If you provide a low rating or 'No' answer to any dimensions, please explain your rating below.
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Panelists reviewed the Science MTAS assessments for Grades 5, 8, and High School and 11 on Criterion 6: Achievement using the following rating form in electronic format. The format of the rating form was identical for each grade assessment.

Evaluation of Student Learning Evident from Scoring Procedures: Grade 3-5

Dimensions	Student Learning	Notes/Comments
	Evidence of Student Learning H-High Student Inference L-Low Student Inference N-No Student Inference	**Rationale for Rating (provide evidence)**
Level of accuracy		
Level of independence		
New learning		
Generalization across people and settings		
Generalization acoress 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: Grade 3-5

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 of 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 the cells 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?										