DFR-08-88

HumRRO

Human Resources Research Organization

Independent Alignment Review of the Science Minnesota Comprehensive Assessment - Series II (MCA-II)

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

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

Prepared under: Contract No:

September 26, 2008

66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314 • 703.549.3611 • fax. 703.549.9025
www.humrro.org

DFR-08-88

HumRRO

Human Resources Research Organization

Independent Alignment Review of the Science Minnesota Comprehensive Assessment - Series II (MCA-II)

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

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

Prepared under: Contract No:

September 26, 2008

66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314 • 703.549.3611 • fax. 703.549.9025
www.humrro.org

EXECUTIVE SUMMARY

Scope of Work

The Minnesota Department of Education (MDE) requested an external independent alignment study (review and analysis) of the Science Minnesota Comprehensive Assessment - Series II (MCA-II) in grades 5, 8, and high school. Specifically, MDE wanted an evaluation of the alignment of the MCA-II for grades 5, 8, and high school to the Minnesota Academic Standards[1] and the newly constructed alternate achievement standards. Minnesota uses the Science MCA-II in the federal and state accountability programs. The Human Resources Research Organization (HumRRO) was awarded a contract to conduct this alignment study, and work began on June 2, 2008.

MDE requested the alignment study in order to meet both state and federal requirements. The federal requirements of the U.S. Department of Education (USDE) stem from the No Child Left Behind (NCLB) Act of 2001. NCLB challenges each state to establish a coherent assessment system based on solid academic standards. This law calls for states to provide independent evidence of the validity of their assessments used to calculate Adequate Yearly Progress (AYP). All states receiving Title I funds must present evidence of establishing a fair and consistent assessment system that is based on rigorous standards, sufficient alignment between standards and assessments and high-quality educational results. States were required to meet these requirements for science by the 2007–2008 academic year.

An alignment review can provide one form of evidence supporting the validity of the state assessment system. Alignment results should demonstrate that the assessments represent the full range of the content standards and that the assessments measure student knowledge in the same manner and at the same level of complexity as specified in the content standards. All aspects of the state assessment system must coincide, including the academic content standards, achievement standards (linked to cut scores), performance level descriptors, and each assessment.

Methodology

Two different types of alignment evaluations were performed for this Minnesota study. These evaluations involved a comparison of: (a) the 2008 Science MCA-II to the Minnesota Academic Standards, and (b) the assessments to the achievement standards. The content alignment evaluation involved a review by current and recently retired Minnesota educators highly familiar with the content standards and the assessment. For the performance alignment, HumRRO compared the difficulty of the assessment items to the established achievement (proficiency) standards and cut scores. The latter review did not involve external panelists.

[1] Minnesota Academic Standards can be found at http://education.state.mn.us/MDE/Academic_Excellence/Academic_Standards/index.html

Review of Content Alignment and Accessibility

For the content alignment review, HumRRO convened panels of Minnesota educators to review the grades 5, 8, and high school Science MCA-II. The review involved two major tasks: (a) matching the science items to grade span Minnesota Academic Standards for Science, and (b) evaluating test quality with respect to students who take the test.

HumRRO developed three review panels with the assistance of MDE and Pearson, Minnesota's current testing contractor. Panelists were recruited by Pearson from their database of Minnesota educators. Every effort was made to produce panels consisting of teachers reflecting the population of students who take the assessments. Panels were convened in facilities procured through MDE. HumRRO directed the actual reviews independently of MDE and Pearson. Each panel included 4–5 reviewers.

To conduct the content alignment review, HumRRO applied the Webb (2005) alignment method. Dr. Norman Webb developed a procedure to evaluate alignment of the assessment to the content standards using four statistics. These statistics indicate how well an assessment covers the content standards in terms of content breadth and depth. The alignment indicators include:

(1) Categorical concurrence – determines the degree of overall content coverage by the assessment for each content strand.

(2) Range-of-knowledge representation – indicates the specific content expectations (e.g., standard, benchmark) assessed within each strand.

(3) Balance-of-knowledge representation – provides a statistical index reflecting the distribution of assessed content within each strand (i.e., how evenly the content is assessed.)

(4) Depth-of-knowledge consistency – compares the cognitive complexity ratings of the items with the complexity ratings of each content standard.

The content reviews also involved a broad examination of test quality that went beyond content alignment. Other facets of test validity are critical as well, such as whether the assessment enables students to demonstrate what they know. For example, are test items free of biases, clear in language, or appropriate for the grade level? Evaluating these aspects of the assessments ensures that the test items are appropriate and accessible to "the widest possible range of students, including students with disabilities and students with limited English proficiency" (NCLB, 2001, Section 200.2(b)(2)). To examine test quality, panelists evaluated the Science MCA-II on several dimensions at the item level and across each grade as a whole.

All assessments should "be designed from the beginning to be accessible and valid with respect to the widest possible range of students, including students with disabilities and students with limited English proficiency" (NCLB, 2001, Section 200.2(b)(2)). The Science MCA-II underwent bias reviews as part of the item

development process; however, review of quality and accessibility from an independent evaluator provides further confirmation of a fair process and assessment.

Review of Performance Alignment

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

Summary of Results

Key Findings and Conclusions

The results of the alignment and quality reviews provide positive support overall for the content validity of the Science MCA-II for each grade (5, 8 and high school) based on several outcomes. First, panelists found that the test items assessed each targeted content strand. Of those benchmarks within the strands, items were distributed rather evenly across these content expectations, although the proportion of benchmarks actually assessed by test items was somewhat limited. Second, panelists considered whether the majority of test items provide reasonable access to the population of students who take the assessments. Most items were judged appropriate, clear in language and free from bias. Finally, the performance alignment review of achievement standards indicated that assessments can discriminate among students in the range of the established achievement levels.

Alignment of Science MCA-II to Minnesota Academic Standards

Table 1 provides summary conclusions on the alignment of the Science MCA-II to the Minnesota Academic Standards per grade tested. The conclusions are based on the following decision criteria (Webb, 2005):

- Fully aligned – assessments align to all content strands (100%);
- Highly aligned – assessments align to the majority of strands (70%–90%);
- Partially aligned – assessments align well to some strands (50%–69%);
- Weakly aligned – assessments align to less than half the strands (below 50%).

Webb's alignment method does not allow for a *single* judgment of overall alignment across the four alignment indicators. However, one can get a sense of overall alignment between the assessments and standards by looking at all of the alignment indicators together.

Table 1. Summary Alignment Outcomes on Each Webb Criterion by Grade Level for Science MCA-II

	Percentage of Strands that Met Webb Criteria			
Grade Assessment	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
5	Fully aligned (100%)	Fully aligned (100%)	Weakly aligned (25%)	Fully aligned (100%)
8	Fully aligned (100%)	Partially aligned (50%)	Weakly aligned (25%)	Fully aligned (100%)
High School	Fully aligned (100%)	Weakly aligned (0%)	Fully aligned (100%)	Fully aligned (100%)

Quality of Science MCA-II Tests

Table 2 presents the summary outcomes on the item quality ratings. The table includes conclusions regarding the quality of the items on each assessment, along with the percentage of items that received favorable ratings. The conclusions are based on the following decision criteria (adapted from Thompson, Johnstone, Anderson, & Miller, 2005):

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

Table 2. Item Quality Ratings for Science MCA-II by Grade

	Percentage of Items with Acceptable Ratings		
Grade	Written Content	Graphics	Overall Item Quality
5	Good (97%)	Acceptable (86%)	Good (94%)
8	Acceptable (83%)	Acceptable (71%)	Acceptable (87%)
High School	Good (94%)	Acceptable (87%)	Good (97%)

The independent item ratings, along with whole test reviews for each grade span group, suggest that the Science MCA-II functions well for the majority of students who take the assessment. A few items on each grade's assessment may require review to enhance clarity in wording or in accompanying graphics and reduce potential bias against particular student groups.

Performance Alignment

Figure 1 demonstrates the results from the review of the High School Science MCA-II with respect to its achievement standards. Test functioning is depicted by a "test characteristic curve" that describes the Item Response Theory-based relationship between achievement and test performance. The cut scores are within the proportion of the curve that shows the strongest relationship between achievement and percentage of items correct. The figure also notes the 2008 percentages of students within each performance category. The assessment is currently functioning most strongly in the region where most students score, and with only 5% in the top category, there is room for the student population to improve over time. Findings for Grade 5 and 8 are similar.

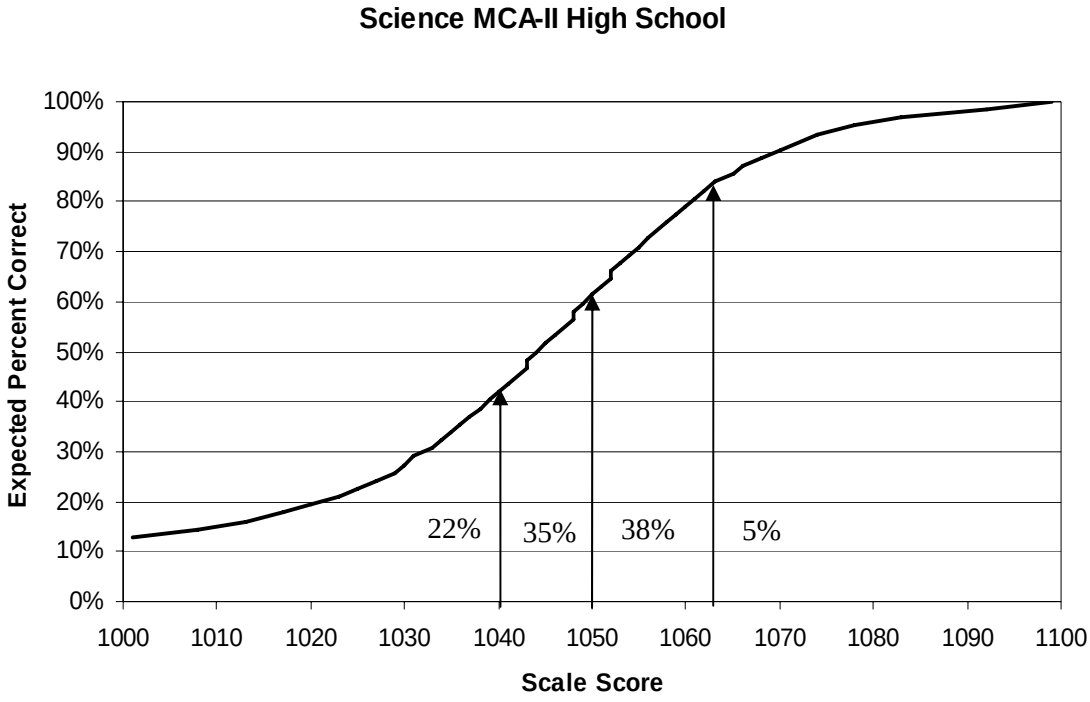

Figure 1. Alignment of achievement levels and High School Science MCA-II test functioning.

Recommendations

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

1. ***Review the cognitive complexity (depth of knowledge) for items on the grade 8 and high school assessments.*** The panelists reviewing these assessments rated a number of items as less cognitively demanding than the Minnesota Academic Standards. Thus, the assessments may not adequately reflect the rigor of the content expectations. Finding a disproportionate

number of items assessing more basic cognitive skills is not uncommon among large-scale assessments. However, such a circumstance is not an inevitable consequence of standardized testing, particularly for dynamic science assessments. Considering the outcomes on depth-of-knowledge ratings, Minnesota may want to consider increasing the complexity of a portion of the assessment items. .

2. ***Examine the extent of content breadth assessed at grades 5 and 8.*** The results of this review indicated that panelists did not match an average of 55% of the grade 5 benchmarks and 61% of the grade 8 benchmarks to items. In other words, substantial portions of the content at those grade levels were not matched to assessment items. Even if the item distribution reflects the intention of the Test Specifications (which is not entirely clear), it is still the case that approximately 25% of benchmarks for grade 5 and 50% of benchmarks for grade 8 cannot be assessed due to the ratio of benchmarks to items. More simply put, there are many more benchmarks to be measured than there are items on a given test form. This finding suggests that the assessments may not adequately "cover the full range of content specified in the State's academic content standards" (USDE, 2004, p.41).

   Two possible ways that Minnesota could address the alignment results to increase the content breadth include the following:

   (a) Although Minnesota has prioritized benchmarks for assessment, it may be worthwhile to review the benchmarks in grades 5 and 8 to determine if the content for some strands could be collapsed to reduce the number of individual content expectations. This approach has been used successfully in other states to reduce granularity.

   (b) Adjust items such that one item is used to assess multiple benchmarks and identify those benchmarks in the Test Specifications. For interactive computer tasks (ICT) science items in particular, this approach may be realistic.

3. ***Review the items that received the lowest ratings on test quality for possible revision.*** As noted in Recommendation 1, no items were rated as seriously flawed or requiring replacement. However, panelists found a small number of items for each grade's test that could benefit from review to increase clarity in language or graphics.

INDEPENDENT ALIGNMENT REVIEW OF THE SCIENCE MINNESOTA COMPREHENSIVE ASSESSMENT - SERIES II (MCA-II)

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

List of Tables

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

List of Figures

INDEPENDENT ALIGNMENT REVIEW OF THE SCIENCE MINNESOTA COMPREHENSIVE ASSESSMENT - SERIES II (MCA-II)

Chapter 1 Introduction

The Minnesota Department of Education (MDE) requested an external independent alignment study of the Science Minnesota Comprehensive Assessment - Series II (MCA-II). Specifically, MDE wanted an evaluation of the alignment of the Science MCA-II for grades 5, 8 and high school to the Minnesota Academic Standards[2] and the newly constructed achievement standards. Minnesota uses the Science MCA-II test in the federal and state accountability programs. The Human Resources Research Organization (HumRRO) was awarded a contract to conduct this alignment study, and work began on June 2, 2008.

MDE requested the alignment study in order to meet both state and federal requirements. The federal requirements of the U.S. Department of Education (USDE) stem from the No Child Left Behind (NCLB) Act of 2001. NCLB challenges each state to establish a coherent assessment system based on solid academic standards. This law calls for states to provide independent evidence of the validity of their assessments used to calculate Adequate Yearly Progress (AYP). All states receiving Title I funds must present evidence of establishing a fair and consistent assessment system that is based on rigorous standards, sufficient alignment between standards and assessments and high-quality educational results. States were required to meet these requirements for science by the 2007-2008 academic year.

An alignment review can provide one form of evidence supporting the validity of the state assessment system. Alignment results should demonstrate that the assessments represent the full range of the content standards and that the assessments measure student knowledge in the same manner and at the same level of complexity as specified in the content standards. All aspects of the state assessment system must coincide, including the academic content standards, achievement standards (linked to cut scores), performance level descriptors and each assessment.

Organization and Contents of the Report

This report contains six chapters. Chapter 2 explains alignment methodologies, including general methods used to evaluate alignment of alternate assessments. Subsequent chapters provide alignment results for comparisons between the components of the assessment system: (a) Chapter 3 presents results of the alignment comparison between the science assessments and the Minnesota Academic Standards; (b) Chapter 4 presents results on the accessibility of the assessments to all students; (c) Chapter 5 includes an analysis of the Science MCA-II tasks against the newly developed alternate

[2] Minnesota Academic Standards can be found at http://education.state.mn.us/MDE/Academic_Excellence/Academic_Standards/index.html

achievement standards; and (d) Chapter 6 provides recommendations for MDE to strengthen the alignment of the Science MCA-II over time.

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

Chapter 2 Alignment Study Design and Methodology

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

Alignment of Assessments and Standards on Content and Performance

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

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

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

These questions essentially can be grouped into two categories—content alignment and performance alignment. However, all alignment evaluations tie back to the state content standards.

Content Alignment and Accessibility

Several methods of alignment are in current use. Most methods involve ratings of several aspects of the assessment items relative to the content standards. The ratings are analyzed statistically to determine the extent of alignment. HumRRO used the alignment method developed by Norman Webb (1997; 1999; 2005) to evaluate the Science MCA-II.

Webb Alignment Method.

The Webb alignment method was originally designed for use with standard large-scale assessments. Dr. Webb has researched and refined this method over time (e.g.,

Webb, 1997; 1999; 2005), and his approach is supported by the Council of Chief State School Officers (CCSSO).

The Webb method includes four major criteria to evaluate alignment. These criteria link with statistical procedures used to assess how well individual portions of the assessments and standards documents actually match. The four alignment criteria are: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation.

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

Depth of Knowledge (DOK) measures the type of cognitive processing required by items and content standards. For example, is a student expected to simply identify or recall basic facts, or is the student expected to use reasoning in manipulating information or strategizing? Using mathematics as an example, a student may be asked to identify the appropriate use of a decimal among several answer choices. This task should be less complex than trying to explain the concept of a decimal and how and why it can be moved.

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

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

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

Performance Alignment and Accessibility

Assessment systems should align to the state academic standards not only in content but also on performance. Performance alignment focuses more on whether student scores are a reflection of what students know and can do. Through a standard-setting process, states determine which scores on the assessment represent various levels of achievement (referred to as achievement, or performance, standards), thus establishing cut scores. The achievement standards must clearly tie to the content standards by identifying the specific content and type of performance expected of students at each level of achievement. Furthermore, the assessment should consist of items that allow for discrimination of student ability at each achievement level. A disconnect between any of these components can lead to inaccurate scores and, consequently, an inaccessible assessment system.

Scope of Alignment Evaluations for Science MCA-II

Two different types of alignment evaluations were performed for this Minnesota study. These evaluations involved a comparison of: (a) the Science MCA-II to the Minnesota Academic Standards, and (b) the assessments to the achievement standards. The content alignment evaluation involved a review by a panel of current and recently retired Minnesota educators highly familiar with the content standards and the assessment. For the performance alignment, HumRRO compared the assessment items to the new achievement descriptors relative to the established achievement (proficiency) standards and cut scores. The latter review did not involve external panelists.

Review of Content Alignment and Accessibility

For the content alignment review, HumRRO convened panels of Minnesota educators to review the grades 5, 8 and high school Science MCA-II. The review involved two major tasks: (a) matching the science items to grade span Minnesota Academic Standards for Science, and (b) evaluating test quality with respect to students who take the test.

Panelists.

HumRRO developed three review panels with the assistance of MDE and Pearson. Panelists were recruited by Pearson from their database of Minnesota educators. Every effort was made to produce panels consisting of teachers reflecting the population of students who take the assessments. Panels were convened in facilities procured through MDE. HumRRO directed the actual reviews independently of MDE and Pearson. Table 2.1 presents the characteristics of the panelists per grade-level of the Science MCA-II.

Table 2.1 Professional and Demographic Characteristics of Science MCA-II Panelists

Professional Position	Number of Panelists	Average Years of Experience [a]	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	Asian/Pacific Islander	American Indian/Alaskan Native
Grade 5													
Teacher	5	28.25 (n = 4)	0	2	0	3	0	5	4	0	1	0	0
Administrator	0		0	0	0	0	0	0	0	0	0	0	0
College Educator	0		0	0	0	0	0	0	0	0	0	0	0
Grade 8													
Teacher	4	19.67 (n = 3)	0	2	2	0	0	4	3	0	1	0	0
Administrator	0		0	0	0	0	0	0	0	0	0	0	0
College Educator	1	12.00 (n = 1)	0	0	1	0	1	0	1	0	0	0	0
High School													
Teacher	3	11.67 (n = 3)	0	1	2	0	1	2	3	0	0	0	0
Administrator	1		0	0	1	0	1	0	0	0	0	0	1
College Educator	0		0	0	0	0	0	0	0	0	0	0	0

[a] No information on experience was available for several panelists; thus, $n < 4$ in this column.

Materials.

Panelists evaluated the alignment of the MCA-II items with the Minnesota Academic Standards using rating forms adapted from Webb (2005). All rating forms were completed electronically in Microsoft Excel. Examples of rating forms and instructions are presented in Appendix C.

Test Forms. Panelists evaluated a single 2008 Science MCA-II test form per grade. Table 2.2 lists the characteristics of the form for the 2008 administration for each grade-level test. Because the test form is a secure document, this report does not include any examples of items or references to specific item content.

Table 2.2 Characteristics of 2008 Science MCA-II Test Forms Reviewed

Grade Level	Total Items per Form	Number of Operational Items	Number of Field Test Items
5	47	36	11
8	54	41	13
High School	67	52	15

The Science MCA-II tests are administered as interactive computer tasks (ICT). Many items include dynamic graphics that demonstrate concepts or require student interaction to formulate a response. Some items present scenarios and response options across consecutive computer screens.

Panelists made most of their content alignment ratings on a print version of the 2008 test form. However, panelists had access to the interactive computer-administered version, as well.

Rating Forms and Instructions. Panelists were given instruction sheets listing the rating tasks and forms, as well as code sheets identifying the range of acceptable codes per task (see Appendix C). Panelists completed two rating forms individually: (a) DOK Ratings of Minnesota Academic Standards, and (b) Item Rating Form. In addition, each grade span group completed a “whole test” rating form through consensus (see Appendix C for samples of each).

Procedures.

HumRRO conducted this alignment review at the Minnesota Department of Education on July 21–22, 2008. The workshops began with introductions of staff and observers. Next, panelists read and signed affidavits of non-disclosure for the secure materials they would be reviewing during the workshop. HumRRO staff gave a presentation describing the purpose of the reviews and alignment research in general. This presentation briefly introduced the alignment tasks the panelists would be performing. Reviewers had the opportunity to practice making ratings during the large group session.

Following the general introduction, panelists began working within their content groups. The Science MCA-II reviewers were split into three groups, one at each grade level (grades 5, 8 and high school). All groups contained five reviewers, except the high school group, which had four reviewers. HumRRO staff supervised each group.

Within their small groups, HumRRO staff further trained reviewers with sample assessment items and by answering questions on rating criteria. Regarding instructions on how to rate standards and items, HumRRO staff provided general suggestions and comments when appropriate; however, they emphasized to reviewers that staff would not give explicit direction on how to rate standards or items because reviewers were valued as content experts. Each panelist received a laptop with rating forms already uploaded and formatted. HumRRO staff provided brief instructions about how to work with the electronic rating forms.

After reviewing sample DOK evaluations as a group, panelists proceeded to rate the benchmarks from the Minnesota Academic Standards relevant to each grade span test. For example, panelists reviewing the grade 5 test rated the benchmarks for grades 3, 4 and 5. Panelists first made independent evaluations without discussion. Once all reviewers had completed their DOK ratings, the groups discussed their ratings to achieve consensus for each benchmark; a voluntary scribe within each group recorded these consensus ratings.

Reviewers then received more specific instructions for rating the items. For training, HumRRO staff facilitated the reviewers in evaluating and discussing sample items as a group. After completing the sample items, reviewers rated the items individually on electronic rating forms on their laptops. Panelists rated the individual items on the 2008 test form for their group on several dimensions, including: (a) content match to the benchmarks in the Minnesota Academic Standards, (b) depth of knowledge required by the item, (c) degree of alignment (i.e., how well the item links to the benchmark), (d) content clarity (i.e., readability), and (e) quality of accompanying graphics (if applicable). Panelists assigned a *primary benchmark* to an item based on a judgment that an item clearly measured this content; however, reviewers could assign an *additional standard* if the item seemed to assess another standard equally to the primary standard. These ratings were conducted individually without consensus.

Finally, panelists worked in their small groups to develop consensus ratings for three additional aspects of the MCA-II tests. HumRRO staff trained the panelists on each task, and then the voluntary scribe from within the small group recorded the group's consensus ratings in pre-formatted Excel spreadsheets. The first consensus task required panelists to rate potential barriers for students in being able to demonstrate knowledge (aspects of the MCA-II as a whole that might prevent students with various disabilities or English learners from fully participating). For the second consensus task, panelists rated the extent to which content differs appropriately across the grade level assessments.

All panelists finished tasks in approximately 1.5 days, although they completed their ratings at different times. Once panelists finished the review, their session ended.

Review of Performance Alignment

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

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

Chapter 3 Results: Content Alignment

In this chapter, we report the results of the content alignment evaluation. These analyses are based on panelists' ratings of the Science MCA-II items

Reliability Results

In this section, we report on two different types of agreement analyses on panelists' ratings. First, we compare panelists' ratings of content match to the test contractor's intended content match. Second, we indicate the inter-rater agreement levels between panelists on the ratings they assigned to tasks for various rating scales. The agreement levels for both types of analyses were sufficiently high as to provide further evidence supporting the validity of the alignment process and outcomes reported here.

Panelist-Test Developer Analyses

Table 3.1 presents the agreement outcomes between panelists and Pearson on the content assessed by items per grade level. Agreement was analyzed at several levels of specificity as shown under the table heading 'Percent Agreement with Pearson Codes'. All of the items were analyzed first for 'Exact Match', which indicates that panelists chose the same strand, substrand, and benchmark for the item as the test developer. If panelists did not show an exact match with Pearson, we determined the percent agreement at the *substrand* level (panelists selected the same strand and substrand as Pearson). Finally, for remaining items, we determined whether panelists at least chose the same *strand* as the test developer. The last column in Table 3.1 shows the percentage of ratings by panelists that did not match the Pearson coding at all on items. Because panelists could assign two content codes to a single item, we counted either code if at least one matched with Pearson. The agreement levels reported in Table 3.1 represent separate analyses; thus, percent agreement in each row adds to greater than 100%.

Table 3.1 Percent Agreement between Panelists and Pearson on Target Content for Operational Items

Grade Level	Number of Operational Items per Form	Total Number of Panelist Ratings across Items	Percent Agreement with Pearson Codes			
			Exact Match	Substrand Match	Strand Match	No Match
5	47	235	94	94	95	5
8	54	264	97	97	99	<1
High School	67	267	90	94	98	2

As Table 3.1 indicates, panelists were highly consistent with Pearson in identifying the assessment target of items even at the most specific (benchmark) content level. Furthermore, panelists differed completely from Pearson on content match for only a few items per grade level test. These findings suggest that the operational science items do, in fact, measure the intended content.

Inter-Rater Reliability

In addition to examining the agreement between the panelists and the test developer, we reviewed how well panelists matched each other on content match.

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

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

Table 3.2 shows the intraclass correlation coefficients, representing the level of agreement among the panelists on the content coding of Science MCA-II items. All coefficients are high, ranging from 0.78 to 0.96.

Table 3.2 Inter-Rater Reliability Coefficients for Content Match

Grade Level Assessment	Number of Panelists	Intraclass Correlation	95% Confidence Interval	
			Lower Bound	Upper Bound
5	5	0.78	0.664	0.869
8	5	0.96	0.942	0.976
High School	4	0.95	0.919	0.963

Table 3.3 presents correlation coefficients for the depth-of-knowledge (DOK) codes assigned by the panelists. These coefficients are also sufficiently high (0.75 to 0.81) to ensure confidence in the judgments of the panelists.

Table 3.3 Inter-Rater Reliability Coefficients for DOK Match

Grade Level	Number of Panelists	Intraclass Correlation	95% Confidence Interval	
			Lower Bound	Upper Bound
5	5	0.79	0.649	0.874
8	5	0.81	0.700	0.880
HS	4	0.75	0.637	0.836

Webb Alignment Results

In this section, we review the general outcomes of item analyses on the four Webb alignment indicators. These analyses only include operational items. More detailed numeric results can be found in Appendix A.

All of Webb's measures begin with calculations for each panelist and build up to a summary of results across raters per content strand. First, we calculated the mean ratings across items for each panelist, and then we determined the mean rating across panelists per strand. Results are presented at the strand level.

Categorical Concurrence. Categorical concurrence describes the extent to which the MCA-II items cover the content strands in the Minnesota Academic Standards for science. Webb recommends a minimum of six test questions to adequately assess each content strand. This criterion serves as a guideline for reasonable content coverage. Table 3.4 summarizes the MCA-II alignment results for categorical concurrence.

Table 3.4 Summary of Categorical Concurrence Results for Science MCA-II

	Mean Number of Items per Strand				
Grade Level	History and Nature of Science	Physical Science	Earth and Space Science	Life Science	Strands with at Least Six Tasks
5	11.20	11.40	11.80	13.60	4 of 4
8	10.60	11.80	14.80	16.00	4 of 4
High School	16.25	NA [a]	NA	50.50	2 of 2

[a] NA = Strands are not taught or assessed at these grades.

As Table 3.4 indicates, the grade 5, 8, and high school assessments all surpassed the minimum requirements for categorical concurrence. The grade 5 and 8 assessments include a sufficient number of items on all four science content strands, and the high school items clearly cover the two strands assessed at the high school level. These results indicate that the Science MCA-II adequately cover the science content students are expected to know across these grade levels.

In addition to identifying the benchmark assessed by each item, we asked panelists to indicate *how well* the item assessed the benchmarks. Panelists rated the extent of item alignment to the benchmarks on a 4-point scale ranging from 'Not aligned to any benchmark' to 'Fully aligned to a benchmark – exemplary item'. Table 3.5 presents the mean number of items (across panelists) at each level of alignment. For each grade assessment, panelists rated items as aligned well to the benchmarks matched to that item.

Table 3.5. Panelist Ratings on Overall Item Alignment

Grade Test	Degree of Alignment	Mean Number of Items per Level	SD	Percent of Items per Level
5	Not at all aligned	0.00	0.00	0
	Weakly aligned	1.67	1.15	4
	Highly aligned	18.20	16.87	40
	Fully aligned	34.75	6.24	74
8	Not at all aligned	0.00	0.00	0
	Weakly aligned	4.80	4.32	9
	Highly aligned	25.40	11.46	47
	Fully aligned	22.80	13.14	42
High School	Not at all aligned	1.00	0.00	1
	Weakly aligned	3.00	1.41	4
	Highly aligned	58.50	9.26	87
	Fully aligned	10.50	12.02	16

Depth-of-Knowledge Consistency. Analyses of depth of knowledge (DOK) measure the type of cognitive processing required of students by content standards. The DOK requirements implied by the benchmarks should be matched by assessment items. To confirm this match, panelists were asked to rate the benchmarks and the science items separately. Webb includes an alignment indicator that directly compares panelists' DOK ratings of content standards and test items, which he refers to as *depth-of-knowledge consistency*.

To make their ratings, panelists used a rating scale (adapted from Webb, 2005) with four levels of cognitive complexity. Further information and examples of the DOK levels are found in Appendix C.

- Level 1 Recognition - simple recall of information (i.e., facts, terms); sequencing; more automatic.
- Level 2 Skills/Concepts - beyond habitual response; applying concepts; problem-solving.
- Level 3 Strategic Thinking - requires basic reasoning, planning, or use of evidence; generating hypotheses.
- Level 4 Extended Thinking - complex reasoning; evaluation of multiple sources or independent pieces of evidence; often over an extended period of time.

Table 3.6 summarizes the depth-of-knowledge consistency results for each grade level of the Science MCA-II. Because reviewers evaluated depth of knowledge at the most specific level of the standards document (benchmarks), the table refers to consistency between the items and the benchmarks to which they were matched. Results are summarized in terms of the percentage of items with cognitive complexity ratings at or above (more complex than) the rating for the corresponding benchmark.

Webb's suggested criterion for this alignment indicator is that at least 50% of the items should have complexity ratings at or above the level of the corresponding benchmark.

Table 3.6 Summary of Depth-of-Knowledge Results for Science MCA-II Operational Items by Grade Level

	Percent of Tasks with DOK At or Above the Level of the Benchmarks per Strand					
Grade Level	History and Nature of Science	Physical Science	Earth and Space Science	Life Science	Number of Strands Assessed Adequately	Specific Strands Assessed Inadequately
5	73	72	67	62	4 of 4	None
8	**45**	56	**37**	56	2 of 4	History and Nature of Science; Earth and Space Science
High School	**42**	NA [a]	NA	**48**	0 of 2	History and Nature of Science, Life Science

[a] NA = Strands not taught or assessed at these grades.

Panelists' ratings on depth-of-knowledge consistency suggest that many of the Science MCA-II items may not assess students at the level expected in the Minnesota Academic Standards. The table indicates that only the grade 5 assessment met the minimum criterion of the Webb method. For grade 8, panelists' ratings using Webb DOK levels imply that less than half of items targeting the History and Nature of Science strand and the Earth and Space Science strand assessed students at the appropriate cognitive complexity. Furthermore, these results suggest that just over half of the grade 8 items assessed students at the appropriate depth expected in the benchmarks for the Physical Science and Life Science strands.

As a result of these outcomes based on the Webb method, we conducted a more in-depth review of panelists' evaluations compared to the assessment targets intended by the test contractor. This analysis required us to map the Webb cognitive levels to the Minnesota cognitive levels. The processing distinctions made by Webb and Minnesota are comparable, and they stem from the same research on Bloom's Taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). However, Minnesota chose to adopt three cognitive levels, whereas Webb makes four distinctions. A comparison of these frameworks suggests that Webb's Level 3 (strategic thinking) and Level 4 (extended thinking) can be collapsed into Cognitive Level C (MCA-II Test Specifications for Science, 2008, p. 5).

The MCA-II Test Specifications for Science include the following table specifying item distributions across cognitive levels per grade test. This table indicates the proportion of items per cognitive level that should be included in each administration (2008, p.14).

Table 3.7 Cognitive Level Distribution of Items in Science from 2008 MCA-II Test Specifications

	Distribution of Items by Cognitive Level		
Grades	Level A	Level B	Level C
5	25-35%	40-50%	20-30%
8	20-30%	40-50%	25-35%
High School	15-25%	40-50%	30-40%

Relative to these proportions, we compared the actual number of items that Pearson assigned to each cognitive level on the 2008 assessments with the mean number of items per cognitive level based on panelists' ratings. Figures 3.1 through 3.3 display the distribution of items by Pearson and the distribution based on panelists' ratings. The x-axis includes the collapsed Webb levels (Level 1, Level 2, and Levels 3/4 together) with the Minnesota cognitive levels (A, B, and C). The y-axis indicates the number of items per level. However, it is important to note that this scale refers to the *mean* number of items for the panelists' distribution, while the scale reflects the actual number of items assigned to each cognitive level for the Pearson item distribution.

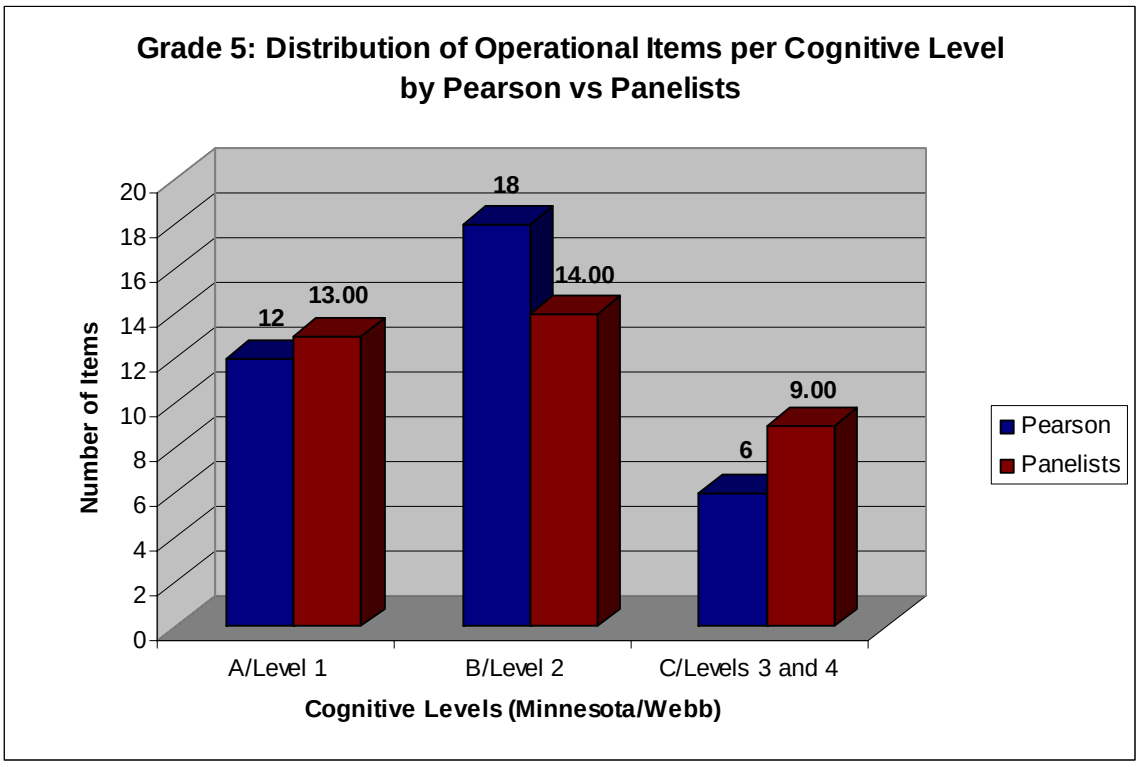

Figure 3.1 Grade 5: Distribution of Items per Cognitive Level based on Pearson Item Assignment compared to Panelists' Mean Item Ratings.

For the grade 5 assessment, panelists' cognitive ratings resulted in a distribution comparable to Pearson's. Thus, the 2008 operational items seem to correspond with the test specifications as intended overall.

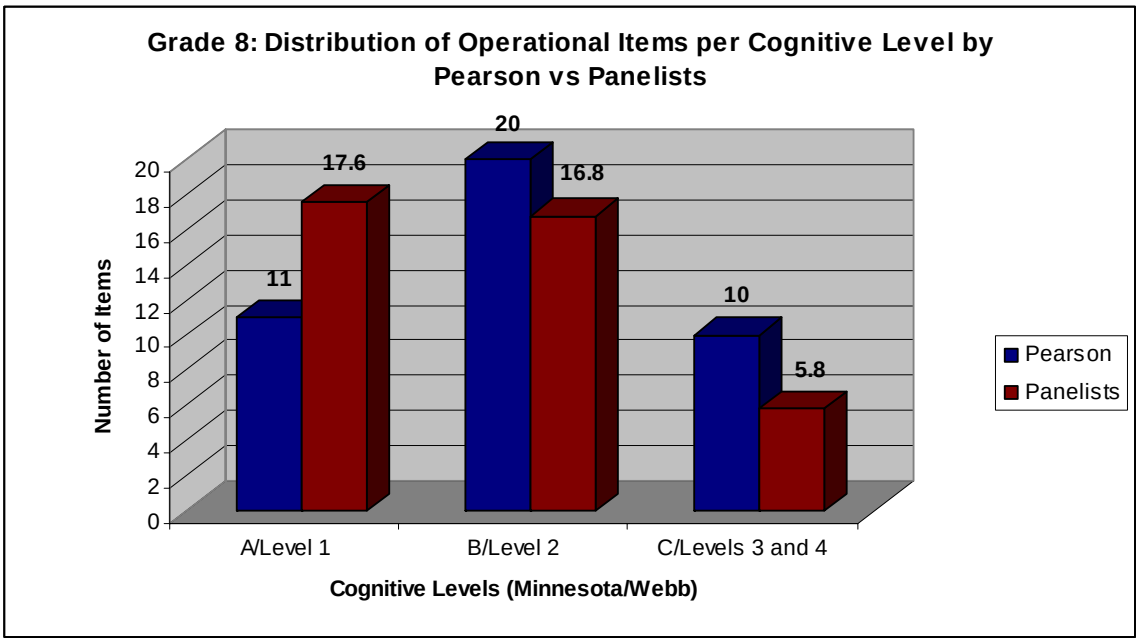

Figure 3.2 Grade 8: Distribution of Items per Cognitive Level based on Pearson Item Assignment compared to Panelists' Mean Item Ratings.

For the grade 8 assessment, panelists' ratings seem to suggest that items tend to be skewed more towards the lower cognitive levels than intended by Pearson. These findings indicate that panelists gave some items lower cognitive complexity ratings than the test developer. Approximately 43% of items fell into Cognitive Level A (instead of the 20-30% intended), and 14% of items corresponded with Cognitive Level C (25-35% expected).

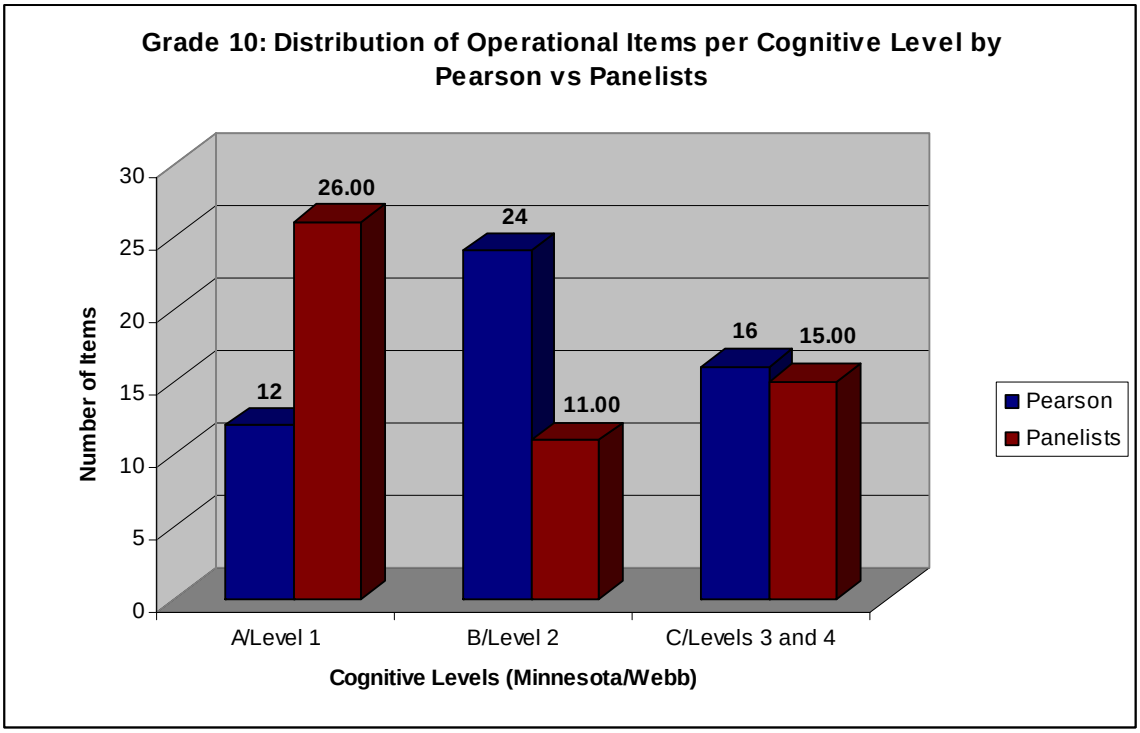

Figure 3.3 High school: Distribution of Items per Cognitive Level based on Pearson Item Assignment compared to Panelists' Mean Item Ratings.

The high school assessment showed the greatest disparity between the test developer and panelists on Cognitive Levels A and B, while MDE and the panelists classified proportionately the same number of items (which matches the test specifications) into Cognitive Level C. Again, panelists' ratings suggest that a larger number of items (M=50% of items classified as Cognitive Level A) than expected require students to demonstrate very basic science knowledge even at the high school level.

Range of Knowledge. The range-of-knowledge measure examines in greater detail the breadth of knowledge covered by the assessment. In addition to evaluating which content strands are assessed, we must look at how many of the benchmarks within a strand are represented by items. The benchmarks should be linked with at least one item. Webb's minimum level of acceptability for range-of-knowledge correspondence is that at least 50% of benchmarks per strand link with items to ensure adequate breadth of content coverage.

Table 3.8 lists the number of strands and benchmarks found in the Minnesota Academic Standards. Some benchmarks are not intended for assessment on the Science MCA-II, as noted (N/A) in the Test Specifications. Column 4 indicates the number of benchmarks that may be represented on the assessment. The last column also indicates the number of items available to assess these strands and benchmarks.

Table 3.8. Number of Content Strands and Benchmarks per Grade Level Science MCA-II

Grade Level Test	Number of Content Strands	Number of Benchmarks per Grade Band	Number of Benchmarks Available for Assessment	Total Items per Form
5	4	64	60	47
8	4	103	98	54
High School	2	50	43	67

To determine how many of these benchmarks were matched to items, we first computed the frequency of benchmarks covered (per strand) separately for each panelist. Next, we calculated the mean number of benchmarks linked with items across panelists. Table 3.9 summarizes the range-of-knowledge results for each grade level of the Science MCA-II per content strand.

Table 3.9. Summary of Range-of-Knowledge Results for Science MCA-II by Grade Level

	Percent of Benchmarks per Strand Matched to at Least One Itemk					
Grade Level	History and Nature of Science	Physical Science	Earth and Space Science	Life Science	Number of Strands Assessed Adequately	Specific Strands Assessed Inadequately
5	**35**	65	**36**	**49**	1 of 4	History and Nature of Science; Earth and Space Science; Life Science
8	**30**	**41**	55	**36**	1 of 4	History and Nature of Science; Physical Science; Life Science
High School	50	NA [a]	NA	81	2 of 2	None

[a] NA = Strands not taught or assessed at these grades.

For each grade assessment, a sufficient number of benchmarks were represented by items for at least one content strand. The high school assessment exhibited adequate range-of-knowledge correspondence for the two content strands included in the high school curriculum. However, as shown in the table, exactly half of the benchmarks for History and Nature of Science were matched to items, which is the minimum number acceptable for adequate representation of the strand.

Panelists for the grade 5 and grade 8 assessments found it difficult to match items to all of the benchmarks for three of four strands. For the grade 5 test, only the Physical Science benchmarks were represented adequately (M=65% of benchmarks matched to at least one item), while just over 50% of the Earth and Space Science benchmarks matched at least one item for grade 8.

Several issues should be considered as explanations for the weak alignment outcomes on range-of-knowledge correspondence. First, because the Science MCA-II is a grade-span test, a substantial amount of content is available for assessment, as demonstrated in Table 3.8. As the number of specific content expectations increase per assessment, the ability of the test to adequately cover these expectations decreases due to practical limits in test length. Second, although Minnesota does prioritize benchmarks for assessment and excludes other benchmarks entirely (i.e., assessed in classroom) in the Test Specifications, the Webb alignment method does not easily allow for weighting of benchmarks. Instead, this approach provides panelists the opportunity to review and match every content expectation in the full academic standards. Thus, panelists were provided with all benchmarks in the Minnesota Academic Standards, including those marked as 'not assessed'. The Webb alignment method takes this approach so that panelists will maintain objectivity and independence in ratings instead of simply confirming the test specifications.

Although the number of benchmarks not targeted for assessment in the Test Specifications is small, HumRRO examined panelists' ratings further to evaluate the impact of making available the non-assessed benchmarks for item ratings. We determined how many times panelists assigned the non-assessed benchmarks to items. As shown in Table 3.10, panelists assigned only one non-assessed benchmark for each grade assessment. In each case, the benchmark came from the History and Nature of Science strand.

Table 3.10. Non-Assessed Benchmarks Assigned to Items by Panelists

Grade	Number of Benchmarks Not Assessed on Science MCA-II	Item Code for Non-Assessed Benchmark Assigned by Panelists
5	4	5.I.B.2
8	5	8.I.A.2
High School	7	9-12.I.C.4

The remaining non-assessed benchmarks were not included in the range-of-knowledge calculations because panelists did not match them to items. Thus, the range analyses represent the benchmarks in the Minnesota Academic Standards available for assessment. Table 3.11 below presents the number of benchmarks per grade assessment (across the grade span and strands) matched to items by panelists.

Table 3.11. Comparison of Benchmarks Matched to Items with Benchmarks Available for Assessment per Grade Science MCA-II

Grade	Number of Benchmarks Available for Assessment	Number of Benchmarks Matched to Items by Panelists	Percentage of Benchmarks Matched to Items by Panelists
5	64	41	64%
8	103	54	48%
High School	50	36	72%

Concerning the prioritization of benchmarks, it is more difficult to evaluate the impact of equal treatment of the benchmarks in a fair and objective way without turning to the test maps for each assessment constructed for the 2008 test forms by Pearson. Although the Test Specifications do suggest numbers of items per benchmark, these specifications include a range of possible items targeting the benchmarks, all of which are relatively comparable. For example, most benchmarks under the grades 3-5 History and Nature of Science strand indicate an expectation of approximately 0 – 3 items for assessment. Generally, we do not compare the results of independent panelists to the test contractor item maps because this procedure would measure alignment to the test contractor, not to the state content standards document or published test blueprint. However, this report includes a complete list of all benchmarks matched to items by panelists, along with the mean number of items matched to each one, in Appendix A. MDE and Pearson may wish to review these results relative to the 2008 test maps.

Our general conclusion regarding the range of content assessed is that MDE should consider developing a strategy that would increase the alignment between the test and academic standards. As specified by the USDE (2004), assessments should align to the content expectations established by the state. Although Minnesota does prioritize the benchmarks (and many benchmarks were matched to test items), it is still the case that a sizeable portion of the benchmarks were not represented by the assessment for each grade test.

Balance-of-Knowledge Representation. The fourth measure of alignment included in the Webb method is *balance-of-knowledge representation.* This measure describes the distribution of items linked to each benchmark within each strand. The number of items should be distributed rather evenly between the benchmarks to achieve good balance.

The content balance is determined by calculating an index, or score, for each strand[3]. 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). An index of 70 or higher suggests that items broadly assess the benchmarks for a strand instead of clustering around one or two benchmarks.

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

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

Table 3.12 summarizes the results on balance-of-content representation per grade for the Science MCA-II. Each grade's assessment surpassed the minimum level of acceptability (index of 70) for demonstrating good content balance among those benchmarks matched to items for each strand.

Table 3.12. Summary of Balance-of-Knowledge Representation Results for Science MCA-II by Grade Level

	Balance Index per Strand					
Grade Level	History and Nature of Science	Physical Science	Earth and Space Science	Life Science	Strands with Adequate Balance	Strands with Limited Balance
5	80	80	70	79	4 of 4	None
8	88	81	81	82	4 of 4	None
High School	80	NA	NA	78	2 of 2	None

These results warrant caution, however. Although the outcomes met the Webb minimum criterion, these results should be examined within the context of the range-of-knowledge outcomes. As shown in Table 3.9 (range), items were matched to a narrow range of benchmarks per strand.

Summary and Discussion on Webb Alignment Indicators

The content alignment review of the Science MCA-II evaluated the operational items compared to the Minnesota Academic Standards on a single 2008 test form for grades 5, 8, and high school. A test form for a given yearly administration should be representative of the full set of items in the pool, and, thus, should align appropriately to the content expectations. Alignment of large-scale assessments to state content standards is a requirement of the No Child Left Behind Act of 2001.

HumRRO applied the Webb alignment method to conduct the review. The overall alignment results for the Science MCA-II were mixed. At each of the three grade levels, the assessments met to the full extent the minimum requirements for at least two of the Webb indicators. Results on other alignment indicators, such as depth-of-knowledge consistency and range-of-knowledge correspondence, suggest that some items represent the benchmarks in a more limited way than expected. We present summary alignment judgments for the Science MCA-II in this section based on the statistical outcomes.

Summary alignment judgments are based on Webb (2005). These summary judgments focus on the percentage of content strands represented well by the assessment. Webb outlined a scale with a range of potential alignment outcomes applied to each of the four indicators:

- Fully aligned – assessments align to all content strands (100%);
- Highly aligned – assessments align to the majority of strands (70%–90%)
- Partially aligned – assessments align well to some strands (50%–69%)
- Weakly aligned – assessments align to less than half the strands (below 50%).

Webb's alignment method does not allow for a *single* judgment of overall alignment across the four alignment indicators. However, one can get a sense of overall alignment between the assessments and standards by looking at all of the alignment indicators together.

Table 3.13 presents the summary alignment outcomes for the Science MCA-II based on the above scale. The table includes a summary judgment for each Webb alignment indicator per grade assessment based on the percentage of strands that met the minimum alignment criteria. This summary table is linked to the bottom row of each of Tables A-1 through A-12 in Appendix A. Thus, these summary judgments reflect a final evaluation of each grade assessment per Webb criteria *across* the strands.

As shown in Table 3.13 with green highlighting, a number of outcomes point to strong content alignment of the Science MCA-II to the Minnesota Academic Standards. Each grade assessment clearly includes a sufficient number of operational items to cover the major content categories (strands), as demonstrated by the outcomes on categorical concurrence. Furthermore, across the grade assessments, the balance-of-knowledge representation results suggest that items seem to be distributed reasonably, at least across benchmarks matched by panelists.

Additional grade specific results were positive as well. For the grade 5 assessment, panelists' DOK ratings on the majority of items corresponded with the DOK levels of the benchmarks. Thus, the grade 5 test assesses student knowledge at the same level of complexity as expected in the content standards. For the grade 8 assessment, panelists found items matching a sufficient number of benchmarks per strand, indicating that the assessment covers reasonable breadth of content.

Some aspects of the assessments demonstrated lower levels of alignment to the content standards on one or more of the Webb criteria. Table 3.13 highlights those results showing partial or weak alignment to the content standards. Yellow highlighting indicates partial alignment to the standards, whereas red highlighting indicates weak alignment to the standards.

Table 3.13. Summary Alignment Outcomes on Each Webb Criterion per Grade Level for Science MCA-II

	Percentage of Strands that Met Webb Criteria			
Grade Assessment	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance-of-Knowledge Representation
5	Fully aligned (100%)	Fully aligned (100%)	Weakly aligned (25%)	Fully aligned (100%)
8	Fully aligned (100%)	Partially aligned (50%)	Weakly aligned (25%)	Fully aligned (100%)
High School	Fully aligned (100%)	Weakly aligned (0%)	Fully aligned (100%)	Fully aligned (100%)

Panelists for the grade 8 and high school assessments found that at least half of items assessed student knowledge at a lower level than expected in the content standards, as evidenced by the depth-of-knowledge consistency results. For grade 8, this outcome occurred for the History and Nature of Science strand and the Earth and Space Science strand. The high school curriculum only includes two strands (Nature and History of Science and Life Science), neither of which were assessed sufficiently at the appropriate cognitive level articulated in the benchmarks based on the ratings of these panelists. These outcomes suggest that the operational items for the grades 8 and high school assessments should be reviewed and modified to meet more fully the cognitive expectations of students in the Minnesota Academic Standards.

The weak levels of alignment for grades 5 and 8 on the range-of-knowledge criterion mostly occurred with the same two strands: History and Nature of Science and Life Science. Additionally, the grade 5 assessment exhibited weak alignment to the benchmarks within the Earth and Space Science strand, while grade 8 exhibited weak alignment to benchmarks within Physical Science. For the most part, these findings can be attributed to the large number of grade span benchmarks for assessment on each test Thus, efforts to include items for *every* benchmark necessarily can be confounded by the relatively few item positions available in a given test administration. This situation reflects a common dilemma faced by states and test contractors in attempting to develop valid assessments, particularly for science due to the NCLB allowance of grade span testing.

Suggestions for improving the alignment between the science assessments and Minnesota Academic Standards are discussed in Chapter 6 Summary and Recommendations.

Chapter 4 Results: Test Quality of Science MCA-II

In this chapter, we report the results of panelists' evaluations of test quality. Alignment of assessments to the state content standards serves as one form of test validity evidence. Other areas of validity are critical as well, such as whether the assessment enables students to demonstrate what they know. For example, are test items free of biases, clear in language, and appropriate for the grade level?

All assessments should "be designed from the beginning to be accessible and valid with respect to the widest possible range of students, including students with disabilities and students with limited English proficiency" (NCLB, 2001, Section 200.2(b)(2)). The Science MCA-II underwent bias reviews as part of the item development process; however, review of quality and accessibility by an independent evaluator provides further evidence of a fair process and assessment. This evaluation of test quality for the Science MCA-II represented a broad review of student access to test content.

An additional reason for a test quality review of the Science MCA-II concerns test administration. The Science MCA-II are administered as interactive computer tasks (ICT). Many items include dynamic graphics that demonstrate concepts or require student interaction to formulate a response. Some items present scenarios and response options across consecutive computer screens. It is important to ensure that no particular class of students is disadvantaged as a result of this administration format.

Panelists evaluated the Science MCA-II on several dimensions at the item level and across each grade test as a whole. Item ratings included review of written content and figures or graphics, and were based on simple yes-no evaluations of item quality. Panelists also made "overall item quality" ratings with annotations to report the rationale for their ratings. Finally, panelists in each grade span group made consensus ratings on specific aspects of the test as a whole. Results reported in this section include those for operational and field test items from the 2008 MCA-II.

Panelists made most of their content alignment ratings based on a print version of the 2008 test form. However, panelists did have access to the interactive computer administered version as well, particularly to evaluate the graphics and presentation format of items.

Written Content.

Panelists rated the language used in the items for the extent to which students of various backgrounds and ability levels could access the science content. Ratings consisted of 'yes' or 'no' responses. Table 4.1 below indicates the mean number of items per grade test rated as accessible or not. As the table demonstrates, the majority of items were rated favorably on accessibility.

Table 4.1 Mean Number of Items Rated As Accessible in Content to Range of Students per Grade Assessment.

	Is item content accessible to the range of students who take the assessment?			
Grade	Yes		No	
	Mean number of items	SD	Mean number of items	SD
5	45.60	1.52	2.33	1.15
8	45.00	2.35	9.00	2.33
High School	63.20	2.58	3.98	2.61

If panelists responded 'no', we asked them to provide an explanation of their responses. Most comments pertained to confusing or complex language in items. Several panelists who reviewed the grades 8 and high school assessments did rate a couple of items as potentially biased against some student groups. A summary of panelist comments is found in Appendix B.

Figures and Graphics.

For those science items accompanied by pictures, figures, or graphs, panelists evaluated whether these graphics would be understandable to a wide range of students from different backgrounds and ability levels. Table 4.2 indicates that panelists' ratings were mostly positive.

Table 4.2 Mean Ratings on Accessibility of Figures or Graphics to Range of Students per Grade Assessment.

	Are item figures or graphics accessible to the range of students who take the assessment?			
Grade	Yes		No	
	Mean number of items	SD	Mean number of items	SD
5	40.60	3.78	3.00	2.00
8	38.20	3.18	5.20	3.27
High School	58.50	3.61	3.75	2.50

For those items with graphics rated as not accessible, panelists' comments focused on whether some graphics could be misleading to students and on whether the graphic was even relevant or helpful. A summary of panelist comments is found in Appendix B.

Overall Item Quality.

In addition to rating items on accessibility, panelists had the opportunity to give items a general rating reflecting their judgments of quality. This rating encompassed aspects such as clarity (e.g., wording or item scenario, prompt, or response options) and appropriateness (e.g., off-grade, exceeds benchmark).

- Poor quality - item exhibits serious flaw; recommend replacement.
- Fair quality - item exhibits minor but repairable flaw.
- Good quality - item exhibits no real flaws and is typical for this type of assessment.
- Exceptional quality - item is exemplary for this type of assessment.

Table 4.3 displays the mean ratings on overall item quality per grade assessment. As the table illustrates, panelists considered the vast majority of items to be 'good' to 'exceptional' in quality.

Table 4.3 Panelist Ratings on Overall Item Quality

Grade Test	Item Quality	Mean Number of Items per Level	SD	Percent of Items per Level
5	Poor	0.20[a]	0.45	2
	Fair	2.40	1.14	5
	Good	21.60	14.74	46
	Exceptional	22.80	15.82	49
8	Poor	0	0	0
	Fair	6.80	3.11	13
	Good	27.60	15.53	51
	Exceptional	18.40	15.50	34
High School	Poor	3.00	2.83	4
	Fair	2.50	0.58	4
	Good	60.25	6.24	90
	Exceptional	5.00	5.66	7

[a] One panelist rated a single item as 'poor quality'.

For those items rated as 'fair' or 'poor' in quality, we asked panelists to provide comments to identify the issue and suggest improvements. Many items falling into these categories received comments regarding clarity or complexity. Notations for other items suggested that, while the item aligned to the benchmark overall, the expectations for students to respond to the item exceeded the content expectations of the benchmarks (i.e., item asked students to 'explain', while benchmark only asks students to 'identify'). A summary of panelist comments is found in Appendix B.

Whole Test Evaluation.

At the end of the review session, after panelists had completed all other independent ratings, each grade-span group reviewed the test as a whole to provide more global perspectives on the ability of students to demonstrate their knowledge on the assessment.

The whole test review included five questions to guide panelists' evaluations. The group was expected to discuss their perspectives based on the independent item

ratings just completed, and then generate written conclusions for each question at a global level. Tables 4.4 through 4.6 presents these questions, along with the consensus responses given by each grade-span group. Overall, panelists were in agreement that the Science MCA-II are accessible and appropriate for Minnesota students. Some comments point to particular features of the assessments that they considered to be particularly positive or negative.

Table 4.4 Grade 5 Science MCA-II: Consensus Ratings on Whole Test Evaluation

Guiding Questions	Overall Evaluation (Yes = mostly to all, No = somewhat to none)	Comments Supporting Ratings
Is the computer administered assessment format effective for this population of students?	Yes	More realistic, interactive, read to the students so reading issues are eliminated. This generation is geared toward technology.
Is language clear and appropriate for a science test?	Yes	Students are tested by multiple modalities, reading, listening. Language used is expected science vocabulary for grade level.
Are graphics used clear and appropriate?	Yes	Hot spots (+) allow students to complete test without frustration. Only use bar graphs. Labels are adequate and clear. Graphics with characteristics of birds are unclear.
Is the level of language proficiency expected by test items appropriate?	Yes	Language is consistent throughout the test. Audio provides tutoring for students who can't read the words.
Is this assessment accessible to all students who will take it?	Yes	As long as students have accessibility to the technology. If a school/district lacks computer technology, technology should be provided.

As a general observation, the grade 8 and high school evaluations resulted in two features worth noting. The grade 8 assessment (Table 4.5) received affirmative responses to each question, but some written comments clearly point to issues that panelists found problematic in specific items. For the high school whole test review, panelists opted to respond to only three of five questions. The panelists were all provided instructions for each of the five questions, but a printing error on the response sheets used by the high school panelists may have contributed to their not reaching consensus on the appropriateness of language items. They were provided space to note any issues with particular items on their response sheets and no comments indicating language issues were recorded.

Table 4.5 Grade 8 Science MCA-II: Consensus Ratings on Whole Test Evaluation

Guiding Questions	Overall Evaluation (Yes = mostly to all, No = somewhat to none)	Comments Supporting Ratings
Is the computer administered assessment format effective for this population of students?	Yes	Novelty value of doing it on the computer is probably still enough of a hook to make students pay attention. Kids engaged and take their time….they did not with the reading and math tests I saw them take. However, as we basically did this as a paper and pencil test, we really feel that it really could be administered as a pencil and paper test without the computer lab scheduling fiascos.
Is language clear and appropriate for a science test?	Yes	Many items were confusing. It really is just a glorified written test - the graphics and scenarios were very disconnected from the questions. There are still too many questions where the information presented in the scenario (1, 14, 15, 21, 24) does not relate to the question asked and in fact may actually distract from the student's ability to activate the appropriate schema needed to answer the question. Many misleading diagrams.
Are graphics used clear and appropriate?	Yes	In addition to the scenarios mentioned above, the following (6, 8, 13) had graphics that were distractors, even if the text was mostly o.k.
Is the level of language proficiency expected by test items appropriate?	Yes	For the most part, it seemed readable and likely understandable by many students. It is not a “reading” test.
Is this assessment accessible to all students who will take it?	Yes	Somewhat. Having the reading and audio both should help many more populations have access. Still may be difficult for ELL and students with disabilities. And, “clogging up” of computer labs is a problem.

Table 4.6 High School Science MCA-II: Consensus Ratings on Whole Test Evaluation

Guiding Questions	Overall Evaluation (Yes = mostly to all, No = somewhat to none)	Comments Supporting Ratings
Is the computer administered assessment format effective for this population of students?	Yes	The delivery options (audio, visual) of the test are a good use of technology and allow for accommodation when necessary, however, does this technology put too little emphasis on reading? The computer made the test "come alive".
Is language clear and appropriate for a science test?	No response	No comments provided.
Are graphics used clear and appropriate?	Yes	Drag and drop is good for students interacting with the test. Item comments - #26 - gel image should be on the same page as the question. #36 - partial tree of life graphic cues answer for item. Media comment - there seemed to be inconsistency with the length of time media played; sometimes the image was still for a good length of time. Audio comment - students should be given the option to choose what parts of the exam, if any, to have read (maybe have a start button for audio on each screen).
Is the level of language proficiency expected by test items appropriate?	No response	No comments provided.
Is this assessment accessible to all students who will take it?	Yes	No comments provided.

Summary and Discussion of Test Quality Results

The results of the test quality review by panelists suggest that the Science MCA-II allow a wide range of students the opportunity to demonstrate their knowledge of science. The majority of items received positive ratings by panelists, and global judgments about test quality also emphasized this point.

Table 4.7 presents the summary outcomes on the item quality ratings. The table includes conclusions regarding the quality of the items on each assessment, along with the percentage of items that received favorable ratings. These conclusions are based

on the following decision criteria (adapted from Thompson, Johnstone, Anderson, & Miller, 2005).

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

Table 4.7. Item Quality Ratings for Science MCA-II per Grade Assessment

	Percentage of Items with Acceptable Ratings		
Grade	Written Content	Graphics	Overall Item Quality
5	Good (97%)	Acceptable (86%)	Good (94%)
8	Acceptable (83%)	Acceptable (71%)	Acceptable (87%)
High School	Good (94%)	Acceptable (87%)	Good (97%)

Table 4.7 shows that none of the grade assessments included enough items with low ratings on any of dimension to warrant a conclusion of questionable quality. However, each assessment included *some* items with low ratings (and corresponding annotations highlighting possible issues), as demonstrated by findings of 'acceptable' quality (70%-90% of items). Panelists for the grade 8 assessment in particular commented on a number of items with graphics that were either unclear or unnecessary (not adding to the item). For this reason, at least those items with low ratings could be reviewed for improvement.

As a whole, the independent item ratings, along with whole test reviews, suggest that the MCA-II function well for the majority of students who take these assessments. A small number of items on each grade assessment may require review to enhance clarity in wording or in accompanying graphics and reduce potential bias against particular student groups.

Chapter 5 Results: Science MCA-II Items and Achievement Standards

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

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

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

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

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

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

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

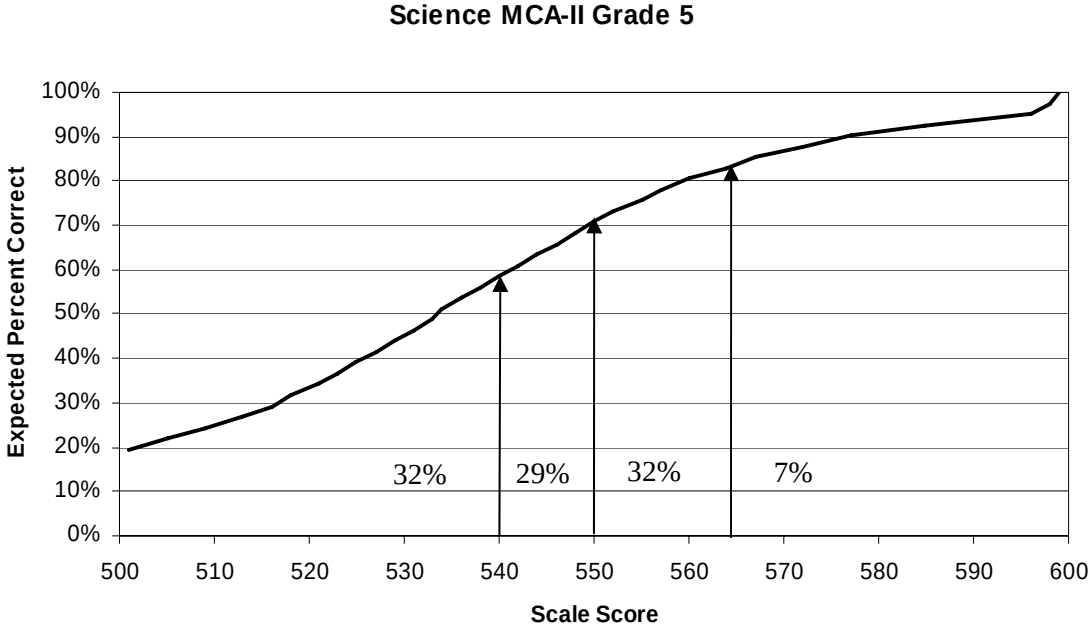

Figure 5-1. Alignment of achievement levels for Grade 5 Science.

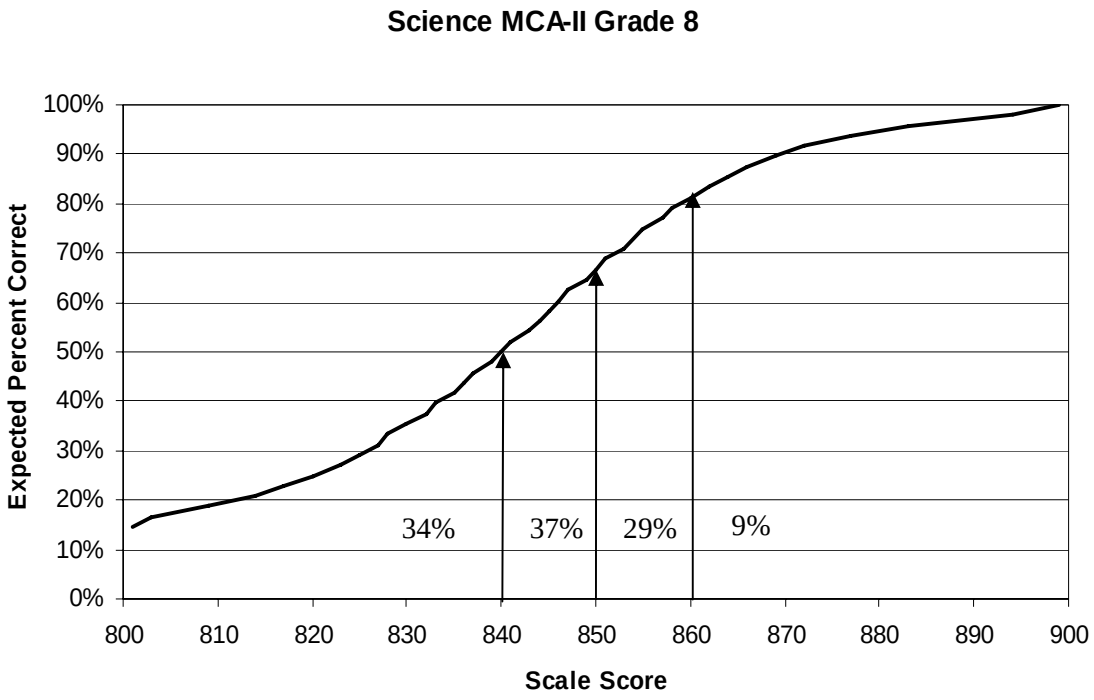

Figure 5-2. Alignment of achievement levels for Grade 8 Science.

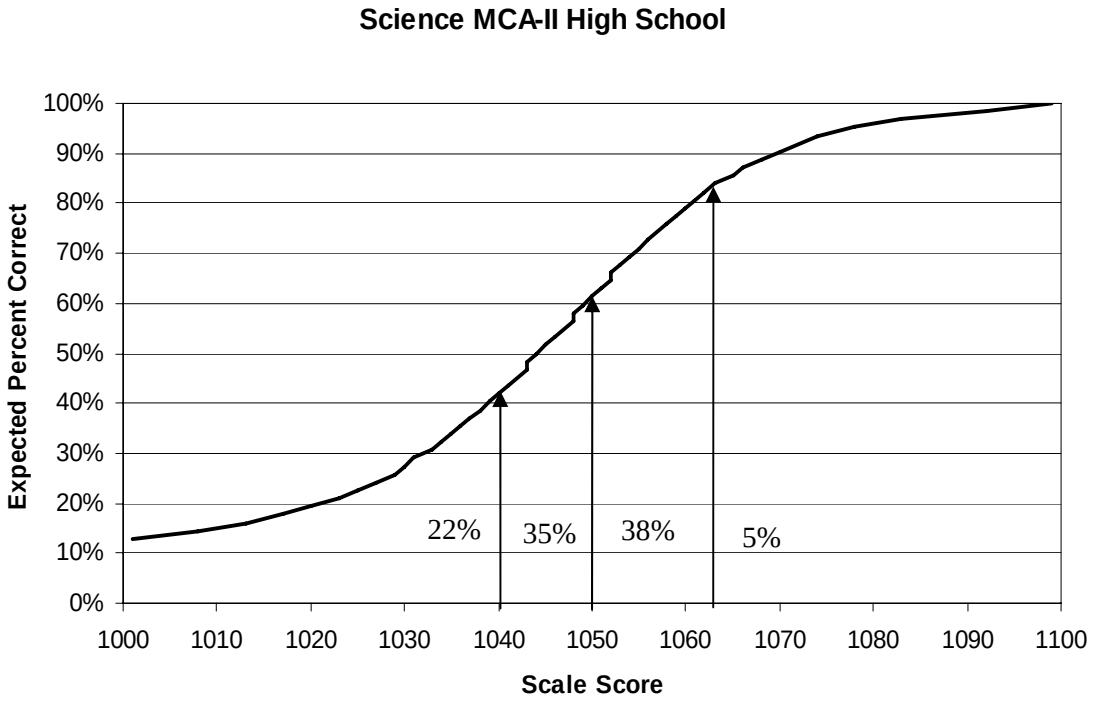

Figure 5-3. Alignment of achievement levels for High School Science.

Summary and Discussion of Science MCA-II Assessment and Achievement Standards

The figures above indicate that the Science MCA-II assessments generally function best in the range of the achievement level cut points. That is, in each case, the steeper parts of the TCCs cover the range of the achievement level cuts. In addition, the Science MCA-II assessments also contain items of varied difficulty such that they cover the range of most students' achievement. Thus, the Science MCA-II assessments meet requirements of alignment of overall test difficulty with the achievement level standards and with the achievement levels of the student population.

Chapter 6 Summary and Recommendations

HumRRO conducted a review of the Science MCA-II to examine: (a) content alignment to the Minnesota Academic Standards for science, (b) performance alignment to the achievement standards, and (c) accessibility for all students who take these assessments. Alignment of assessments and achievement standards to the state academic content standards is a requirement of the No Child Left Behind Act of 2001.

The cumulative results provide reasonable evidence for content validity of the Science MCA-II. Concerning content alignment, each assessment clearly covers the content categories specified in the Minnesota Academic Standards for science. Concerning accessibility, panelists determined that the majority of items are appropriate for a wide range of students. Finally, the achievement standards corresponding with the Science MCA-II appear to categorize students in an accurate and consistent manner, which lends support to appropriate alignment of performance expectations with the assessment and content standards.

As with most reviews of state assessment systems, these findings point to areas where Minnesota could strengthen the alignment between the assessments and the content standards. For this reason, HumRRO makes the following recommendations to Minnesota on ways in which alignment might be improved. These recommendations focus on the more critical findings.

Recommendations:

1. ***Review the cognitive complexity (depth-of-knowledge) for items on the Grades 8 and high school assessments.*** The panelists reviewing these assessments rated a number of items as less demanding cognitively than the Minnesota Academic Standards. Thus, the assessments may not adequately reflect the rigor of the content expectations. Finding a disproportionate number of items assessing more basic cognitive skills is not uncommon among large-scale assessments. However, such a circumstance also is not an inevitable consequence of standardized testing, particularly for dynamic science assessments. Given the outcomes on depth-of-knowledge ratings along with the accessibility outcomes, it is likely that increasing the complexity of the assessment would involve modifications to current operational items, rather than item replacement, because no items were rated as seriously flawed.

2. ***Examine the extent of content breadth assessed at Grades 5 and 8.*** The results of this review indicated that panelists could not match as many as 35% of the grade 5 benchmarks and 55% of the grade 8 benchmarks to items. Even if the item distribution reflects the intention of the Test Specifications (which is not entirely clear), it is still the case that approximately 25% of benchmarks for grade 5 and 50% of benchmarks for

grade 8 cannot be assessed due to the ratio of benchmarks to items. This circumstance suggests that the assessments may not adequately "cover the full range of content specified in the State's academic content standards" (USDE, 2004, p.41).

Two possible ways that Minnesota could address the alignment results to increase the content breadth include:

(a) Although Minnesota has prioritized benchmarks for assessment, it may be worthwhile to review the benchmarks in grades 5 and 8 to determine if the content for some strands could be collapsed to reduce the number of individual content expectations. This approach has been used successfully in other states to reduce granularity.

(b) Adjust items to assess multiple benchmarks and identify in the Test Specifications. For ICT science items in particular, this approach should be realistic.

3. ***Review those items that received the lowest ratings on test quality for possible revision.*** As noted in Recommendation 1, no items were rated as seriously flawed or requiring replacement. However, panelists did find a small number of items for each grade test that could benefit from review to increase clarity in language or graphics.

References

Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, Toronto: Longmans, Green.

No Child Left Behind Act of 2001. Public Law 107-110.

Thompson, S.J., Johnstone, C.J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (April, 2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001.* Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education. Retrieved from http://www.ed.gov/policy/elsec/guid/saaprguidance.doc.

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

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

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

Appendix A
Content Alignment Results per Grade Level Assessment

The following tables include complete statistical results on the Webb alignment indicators, including means and standard deviations per strand for each grade Science MCA-II test.

Categorical Concurrence

The categorical concurrence results for grades 5, 8 and high school of the Science MCA-II are presented below. Each table includes: the target number of items from the test blueprint; the mean number of items matched by panelists; the standard deviation among panelists' ratings; and, the final alignment conclusion (Yes or No). The bottom row indicates the percentage of strands that met the minimum alignment criterion. Note that the total mean tasks matched may exceed the number of items on the assessment, as raters were able to match items to more than one strand.

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

	Number of Tasks per Strand			
Title of Strand	Target # Items from Blueprint	Mean Tasks Matched	Standard Deviation	At Least One Task per Strand
History and Nature of Science	9-11	11.20	1.10	Y
Physical Science	8-10	11.40	2.07	Y
Earth and Space Science	8-10	10.80	2.39	Y
Spatial Sense, Measurement, and Geometry	8-10	13.60	0.55	Y
Total	33-37	47.00		
Percent of strands with at least one task				100%

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

	Number of Tasks per Strand			
Title of Strand	Target # Items from Blueprint	Mean Tasks Matched	Standard Deviation	At Least One Task per Strand
History and Nature of Science	8-10	10.60	1.52	Y
Physical Science	10-12	11.80	0.84	Y
Earth and Space Science	10-12	14.80	1.30	Y
Spatial Sense, Measurement, and Geometry	10-12	16.00	1.22	Y
Total	38-42	53.20		
Percent of strands with at least one task				100%

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

Title of Strand	Number of Tasks per Strand			At Least One Task per Strand
	Target # Items from Blueprint	Mean Tasks Matched	Standard Deviation	
History and Nature of Science	15-17	16.25	0.96	Y
Physical Science	NA	NA	NA	NA
Earth and Space Science	NA	NA	NA	NA
Spatial Sense, Measurement, and Geometry	35-37	50.50	0.58	Y
Total	48-52	66.75		
Percent 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 MCA-II are presented below. The tables present the results from the comparison between the depth-of-knowledge expected in the content benchmarks and the depth-of-knowledge assessed by items. The tables include the mean percentage of items rated as below, at the same level, or above the DOK level of the benchmarks along with the corresponding standard deviations. Benchmarks with at least 50% of items at the same (or above) DOK level met the minimum criterion.

Table A-4. DOK Consistency for Science MCA-II, Grade 5: Mean Percent of Items with DOK Below, At, and Above DOK Level of Benchmarks

Title of Strand	Mean Tasks per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
History and Nature of Science	11.20	27	8.91	48	11.19	25	14.08	Y
Physical Science	11.40	29	15.33	51	12.39	21	16.50	Y
Earth and Space Science	10.80	33	21.07	54	14.35	13	20.95	Y
Life Science	13.60	38	16.25	43	11.97	19	15.29	Y
Percent of strands with 50% of item DOK at or above objective DOK:								100%

Table A-5. DOK Consistency for Science MCA-II, Grade 8: Mean Percent of Items with DOK Below, At, and Above DOK Level of Benchmarks

Title of Strand	Mean Tasks per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
History and Nature of Science	10.60	55	15.38	33	15.38	12	3.85	N
Physical Science	11.80	43	16.54	41	14.52	15	6.69	Y
Earth and Space Science	14.80	64	11.15	28	5.09	9	8.61	N
Life Science	16.00	44	10.09	46	9.03	10	7.01	Y
Percent of strands with 50% of item DOK at or above objective DOK:								50%

Table A-6. DOK Consistency for Science MCA-II, High school: Mean Percent of Items with DOK Below, At, and Above DOK Level of Benchmarks

Title of Strand	Mean Tasks per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Tasks Below		% Tasks Same Level		% Tasks Above		
		M	S.D.	M	S.D.	M	S.D.	
History and Nature of Science	16.25	58	22.31	28	15.94	14	8.82	N
Physical Science	NA	NA	NA	NA	NA	NA	NA	NA
Earth and Space Science	NA	NA	NA	NA	NA	NA	NA	NA
Life Science	50.50	52	9.25	35	1.54	13	9.44	N
Percent of strands with 50% of item DOK at or above objective DOK:								0%

Range-of-Knowledge Correspondence

The results for Range-of-Knowledge correspondence for grades 5, 8 and high school for the Science MCA-II are presented below. The tables include the mean number, standard deviation, and percentage of benchmarks by content strand. For acceptable range-of-knowledge correspondence, a minimum of 50% of content benchmarks within each strand should be matched to at least one item.

Table A-7. Range-of-Knowledge for Science MCA-II, Grade 5: Mean Percent of Benchmarks per Strand Linked with Items

Title of Strand	Number of Benchmarks	Mean Tasks per Strand	Range of Benchmarks			Range-of-Knowledge Target Met
			Benchmarks with At Least One Task		% of Total Benchmarks per Strand	
			M	S.D.	M	
History and Nature of Science	16	11.20	5.60	1.14	35	N
Physical Science	12	11.40	7.80	1.10	65	Y
Earth and Space Science	17	10.80	6.20	0.45	36	N
Life Science	19	13.60	9.40	1.95	49	N
Total	64	47.00				
Percentage of strands with 50% of Benchmarks linked to at least one item						25%

Table A-8. Range-of-Knowledge for Science MCA-II, Grade 8: Mean Percent of Benchmarks per Strand Linked with Items

Title of Strand	Number of Benchmarks	Mean Tasks per Strand	Range of Benchmarks			Range-of-Knowledge Target Met
			Benchmarks with At Least One Task		% of Total Benchmarks per Strand	
			M	S.D.	M	
History and Nature of Science	29	10.60	8.80	1.30	30	N
Physical Science	21	11.80	8.60	1.14	41	N
Earth and Space Science	19	14.80	10.40	0.89	55	Y
Life Science	34	16.00	12.20	1.48	36	N
Total	103	53.20				
Percentage of strands with 50% of Benchmarks linked to at least one item						25%

Table A-9. Range-of-Knowledge for Science MCA-II, High school: Mean Percent of Benchmarks per Strand Linked with Items

Title of Strand	Number of Benchmarks	Mean Tasks per Strand	Range of Benchmarks			Range-of-Knowledge Target Met
			Benchmarks with At Least One Task		% of Total Benchmarks per Strand	
			M	S.D.	M	
History and Nature of Science	19	16.25	9.50	1.29	50	Y
Physical Science	NA	NA	NA	NA	NA	NA
Earth and Space Science	NA	NA	NA	NA	NA	NA
Life Science	31	50.50	25.25	0.96	81	Y
Total	50	66.75				
Percentage of strands with 50% of Benchmarks linked to at least one item						100%

Balance-of-Knowledge Representation

The results for Balance-of-Knowledge representation for grades 5, 8 and high school of the Science MCA-II are presented below. The tables also include the percentage of items linked to each strand. The minimum acceptable balance index is 70 out of 100.

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

Title of Strand	Balance-of-Knowledge Representation						
	Benchmarks per Strand	Mean 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.	
History and Nature of Science	16	5.60	11.20	24	80	3.94	Y
Physical Science	12	7.80	11.40	24	80	3.68	Y
Earth and Space Science	17	6.20	10.80	23	70	7.03	Y
Life Science	19	9.40	13.60	29	79	6.63	Y
Total	64	41.4	47.00				
Percentage of standards with a balance of representation index of 70 or greater							100%

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

Title of Strand	Balance-of-Knowledge Representation						
	Benchmarks per Strand	Mean 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.	
History and Nature of Science	29	8.80	10.60	20	88	5.47	Y
Physical Science	21	8.60	11.80	22	81	8.40	Y
Earth and Space Science	19	10.40	14.80	28	81	3.15	Y
Life Science	34	12.20	16.00	30	82	2.99	Y
Total	103	40.00	53.20				
Percentage of standards with a balance of representation index of 70 or greater							100%

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

Title of Strand	Balance-of-Knowledge Representation						
	Benchmarks per Strand	Mean 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.	
History and Nature of Science	19	9.50	16.25	24	80	2.17	Y
Physical Science	NA	NA	NA	NA	NA	NA	NA
Earth and Space Science	NA	NA	NA	NA	NA	NA	NA
Life Science	31	25.25	50.50	76	78	0.61	Y
Total	50	34.75	66.75				
Percentage of standards with a balance of representation index of 70 or greater							100%

Tables A-13 through A-15 present the benchmarks, along with mean number of items, matched by panelists. Column 1 includes the Item Codes corresponding to the benchmarks from the MCA-II Test Specifications for Science.

Table A-13. Grade 5 MCA-II: Grade Span Benchmarks Matched to Items by Panelists

Benchmark Item Codes	Mean Number of Items per Benchmark	SD
3IA1	2.00	0.00
3IB1	2.00	1.00
3IB2	3.00	0.00
3IB3	1.00	0.00
3IIC1	1.00	0.00
3IIC2	1.00	0.00
3IIIB1	1.33	0.58
3IIIB2	0.20	0.45
3IIIC1	0.20	0.45
3IIIC2	1.20	0.45
3IVB1	1.33	0.58
3IVB21	0.20	0.45
3IVC1	1.25	0.50
3IVC2	2.00	1.41
3IVD1	1.00	0.00
4IB1	1.20	0.45
4IB2	2.20	0.84
4IB3	2.00	1.00
4IIA1	2.75	0.96
4IIA2	1.50	0.58
4IIA3	1.00	0.00
4IIC1	1.60	0.55
4IIC3	1.60	0.89
4IIE1	1.00	0.00
4IIE2	1.75	0.50
4IIIA1	0.20	0.45
4IIIB1	5.00	1.41
4IIIB2	1.00	0.00
4IIIC1	1.00	0.00
4IVA1	1.00	0.00
4IVA2	0.20	0.45
4IVB1	2.60	1.95
4IVB2	2.00	1.00
4IVG2	0.20	0.45
5IA2	0.20	0.45
5IB1	2.00	1.41
5IB2	0.80	1.79
5IID1	1.00	0.00
5IIIA3	1.20	0.45
5IIIA4	0.20	0.45
5IVE1	1.00	0.00
5IVE2	1.67	1.15
5IVF1	1.50	1.00
5IVF2	1.00	0.00
5IVF3	1.00	0.00

Table A-14. Grade 8 MCA-II: Grade Span Benchmarks Matched to Items by Panelists

Benchmark Item Codes	Mean Number of Items per Benchmark	SD
6IA2	0.20	0.45
6IB2	1.80	0.84
6IB4	1.00	0.00
6IC2	1.00	0.00
6IIA1	1.00	0.00
6IIA2	1.00	0.00
6IIA3	1.20	0.45
6IIA4	0.20	0.45
6IIA6	1.00	0.00
6IIB1	1.00	0.00
6IIB2	0.20	0.45
6IIB3	1.00	0.00
6IIC1	1.80	0.45
6IIC2	3.20	1.30
6IIC5	1.00	0.00
6IID3	1.00	0.00
7IA1	1.00	0.00
7IA2	1.67	0.58
7IB1	0.20	0.45
7IB2	1.00	0.00
7ID1	1.00	0.00
7ID2	1.00	0.00
7IVA3	1.20	0.45
7IVA4	0.20	0.45
7IVA5	1.00	0.00
7IVA6	1.33	0.58
7IVB2	1.67	0.58
7IVB3	0.20	0.45
7IVB5	0.20	0.45
7IVC1	1.25	0.50
7IVC2	1.80	0.45
7IVC4	1.00	0.00
7IVD5	0.20	0.45
7IVE1	1.00	0.00
7IVE3	3.50	1.29
7IVE4	1.00	0.00
7IVF2	1.00	0.00
7IVF3	1.00	0.00
7IVF4	0.20	0.45
7IVF5	0.20	0.45
7IVG1	1.00	0.00
7IVG2	0.20	0.45
7IVG3	2.00	1.15
8IA2	1.25	0.50
8IB2	0.20	0.45
8IB3	1.50	0.71
8IB4	1.00	0.00
8IC1	1.00	0.00
8ID2	0.20	0.45
8IIIA1	1.80	0.84
8IIIA2	1.80	1.30
8IIIA3	1.80	0.45
8IIIA4	1.60	0.55
8IIIA5	0.20	0.45

Benchmark Item Codes	Mean Number of Items per Benchmark	SD
8IIIA6	1.00	0.00
8IIIA7	0.20	0.45
8IIIB1	1.50	0.58
8IIIB2	1.25	0.50
8IIIB4	1.00	0.00
8IIIB5	0.20	0.45
8IIIB6	1.00	0.00
8IIIC1	1.00	0.00
8IIIC2	1.60	0.89
8IIIC4	1.00	0.00

Table A-15. High School MCA-II: Grade Span Benchmarks Matched to Items by Panelists

Benchmark Item Codes	Mean Number of Items per Benchmark	SD
9IA1	2.25	0.50
9IA2	1.25	0.50
9IA5	1.50	1.00
9IB1	2.75	0.96
9IB2	2.50	0.58
9IB3	2.00	1.15
9IB4	1.25	0.50
9IB6	1.00	0.00
9IC2	1.00	0.00
9IC4	1.00	0.00
9ID1	1.33	0.58
9IVA1	3.00	0.82
9IVA2	2.00	0.00
9IVA3	1.00	0.00
9IVA4	1.75	0.96
9IVA5	2.25	0.96
9IVA6	2.00	0.82
9IVB1	1.33	0.58
9IVB3	2.25	0.50
9IVC1	3.25	2.06
9IVC2	2.75	0.96
9IVC3	1.50	1.00
9IVC3	4.00	1.63
9IVD1	3.25	0.50
9IVD2	1.00	0.00
9IVD3	1.33	0.58
9IVD4	1.00	0.00
9IVD5	1.00	0.00
9IVD6	1.33	0.58
9IVD7	1.25	0.50
9IVE1	2.00	0.00
9IVE2	1.25	0.50
9IVE3	1.00	0.00
9IVF1	2.67	1.53
9IVF2	1.50	0.71
9IVF3	1.75	0.96
9IVF4	1.50	1.00
9IVF5	0.25	0.50
9IVG1	2.75	0.96
9IVG2	4.00	0.82

Appendix B
Summary of Panelist Comments on Items

Tables B -1 through B – 3 present a synopsis of panelists' comments on the individual items of the Science MCA-II. To maintain test security, individual item identifiers are not presented, nor are any comments that would reveal the content of a task. Column 3 indicates the number of items receiving such comments, and Column 4 reports how many panelists included this type of comment.

Table B - 1. Grade 5 Science MCA-II: Summary of Panelists' (N=5) Comments on Items by Topic

Comment	Number of items with comment	Number of panelists with comment
• Some features of graphics are difficult to make out/ambiguous.	4	5
• Wording of response options is confusing or misleading.	2	2
• Item scenario is unclear or misleading.	3	3
• Target of assessment is unclear.	2	3
• Item requires knowledge beyond the benchmark expectations.	1	2
• Students can answer question without knowing content.	1	1
• Related items/graphics in block include contradictory information.	1	2

Table B - 2. Grade 8 Science MCA-II: Summary of Panelists' (N=5) Comments on Items by Topic

Comment	Number of items with comment	Number of panelists with comment
• Some features of graphics are difficult to make out/ambiguous.	8	5
• Wording of item is unnecessarily complex.	2	2
• Wording of response options is confusing or misleading.	2	2
• Item scenario is unclear or misleading.	8	5
• Target of assessment is unclear.	10	5
• Item requires knowledge beyond the benchmark expectations.	1	2
• Content could be split into several items.	2	3
• Item assesses non-science skills (e.g., reading comprehension).	1	1
• Graphic and/or item content may be offensive.	1	1
• Related items/graphics in block seem out of sequence.	1	1
• Content of item is off-grade.	3	3

Table B - 3. High School Science MCA-II: Summary of Panelists' (N=4) Comments on Items by Topic

Comment	Number of items with comment	Number of panelists with comment
• Some features of graphics are difficult to make out/ambiguous.	6	1
• Wording of response options is confusing or misleading.	4	3
• Item scenario is unclear or misleading.	6	4
• Graphic and/or item content may be offensive.	3	2
• Related items/graphics in block seem out of sequence.	1	1
• Content of item is off-grade.	3	1
• Target of assessment is unclear.	5	3
• Item requires knowledge beyond the benchmark expectations.	6	2

Appendix C
Sample Alignment Review Materials

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

Science MCA-II
Panelist Instructions

	Rating Task	Documents Needed	File Format
1	DOK of MN Academic Standards	(1) Minnesota Academic Standards for Science (HumRRO Coded)	Print Copy
		(2) DOK Codes for Science	Print copy
2	Science MCA-II Items	(1) Minnesota Academic Standards for Science (HumRRO Coded) – grade spans	Print copy
		(2) DOK Codes for Science	Print copy
		(3) MCA-II items (printed)	Print copy
		(4) MCA-II items (computer administered)	Electronic
		(5) MCA-Sci_ItemRatingForm	Excel spreadsheet
3	Whole Test	(3) MCA-II items (printed)	Print copy
		(4) MCA-II items (computer administered)	Electronic

1 Rate DOK of Minnesota Academic Standards

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

2 Rate Science MCA-II items on multiple dimensions

Open the file 'MCA-Sci_ItemRatingForm'. Rename the file using the naming conventions. Click on the worksheet with the appropriate grade level.

A **Item DOK.** Assign a depth-of-knowledge rating to each item using the same DOK codes. Rate each item on the degree of cognitive processing required of students to answer the item adequately. Enter the DOK level (number) in the spreadsheet under the column labeled Item DOK Rating next to each item number.

B **Standards Match.** Use the 'MN Academic Standards with HumRRO Codes' to identify the benchmark that the item targets using the numeric code found in the right-hand column.

C **Degree of Alignment.** Rate the overall match level of the item to the benchmark to Indicate *how well* you think that the item actually links to listed benchmark. Using the rating scale below, enter the appropriate rating number from the scale into your spreadsheet under the column 'Overall Alignment'.

1 Not aligned to any benchmark (Use ONLY if you did not assign a benchmark to the item).

2 Weakly aligned to this benchmark – does not assess the content of the academic standards well.

3 Highly aligned to this benchmark - targets core content reasonably well.

4 Fully aligned to the benchmarks - Exemplary item, clear example of standard to which it is matched.

D **Item Quality**. Rate the overall quality of the item. Is the item clear and precise? Could you understand what the item is asking students to do (NOT whether you are capable of answering the item correctly)? Use the scale below to make your judgments.

Overall Item Quality

1 Item is of poor overall quality (Rating requires annotation).
2 Item is of good quality, but has some easily repairable flaw (Rating requires annotation).
3 Item is of good quality, typical of what you would expect on this and similar tests.
4 Item is of exceptional quality (annotations encouraged).

E **Notes/Comments.** Provide annotations for any item that you give a low rating on degree of alignment (rating of 1 or 2) or on item quality (rating of 1).

This rating task will occur at the end of Day 2. Only a few panelists may have time to complete these ratings, depending on time.

4 Rate 'Whole Test' barriers to demonstrating student knowledge

Open the Excel 'MCA-Sci_WholeTestRatings' file. Click on the appropriate grade worksheet.

Please evaluate BOTH the written and the electronic versions of the test to make these ratings. 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.

Panelists received the following coding sheet as a reference guide for the DOK rating scale.

Depth-of-Knowledge (DOK) Levels for Science

(adapted from *Web Alignment Tool (WAT) Training Manual*)

- **Level 1 (recall/reproduction)** item requires recall of information such as fact, definition, term or simple procedure as well as performance of a simple science process or procedure.

  Keywords: Identify, define, determine, perform (simple procedure), list.

- **Level 2 (skill/concept)** Item calls for engagement of some mental processing beyond a habitual response. Students required to make some decisions as to how to approach a problem or activity, such as selecting procedures, describing or giving examples of science concepts, deciding how to display or interpret data.

  Keywords: Describe, observe, classify, confirm, organize, distinguish

- **Level 3 (strategic thinking)** Items require students to use reasoning and evidence, plan, and make conjectures. Students should be able to explain phenomena in terms of scientific concepts, explain simple relationships, explain their own thinking and conclusions, solve non-routine problems, and develop research questions.

  Keywords: Connect, explain, analyze, outline procedures, make conclusions, interpret.

- **Level 4 (extended thinking)** Items require student to use complex and abstract reasoning and thinking, often over an extended period of time. Students must design and plan experimental studies, select and appropriate method among alternatives, or deduct the relationship among several variables.

  NOTE: Many on-demand assessment instruments will not include assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated so as to expect students to perform thinking at this level. On-demand assessments that do include tasks, products, or extended responses would be classified as Level 4 when the task or response requires evidence that the cognitive requirements have been met.

  Keywords: Design, plan, and develop experiments; make inferences from results; critique; predict; explain (complex) relationships or differences among variables.

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 3 as an example.

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

Panelists received the Minnesota Academic Standards for science in a rating form in which to make DOK ratings for each benchmark. Panelists entered DOK ratings (1, 2, 3, or 4) in the last column of the table next to each benchmark. The content of the standards was extracted exactly from the full Minnesota Academic Standards document. Only a portion of the standards is replicated for grade 3 as an example.

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

Panelists reviewed the individual Science MCA-II items using the following rating form in electronic format. The format of the rating form was identical for each grade span. The number of items listed per rating form did differ for each grade test.

Item Number	Depth Of Knowledge	Benchmark 1	Benchmark 2	Written Content	Figures/Graphics	Overall Alignment	Overall Item Quality	Explanation
(Number Listed in Test Form)	1-Recall 2-Skill 3-Reasoning 4-Inference	(Enter Standard ID Code)	(Enter Standard ID Code)	Y=universal, clear, unbiased N=needs revision	Y=universal, unambiguous, informative N=needs revision	(Enter Scale of 1 to 4)	(Enter Scale of 1 to 4)	Please provide if you entered an *Overall Alignment rating of '1' or '2'* and/or an *Overall Item Quality* rating of '1'
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								