An Alignment Study of the Minnesota Comprehensive Assessment-II with State Standards in Mathematics for Grades 3-8 and 11

Prepared for the
Minnesota Department of Education
Division of Statewide Assessment & Testing
Under Contract #A87623

by

Thomas J. Lombard, Ph.D.
651-465-4063
tl1946@hotmail.com

September 1, 2006

Minnesota Department of Education
11/8/2007

Abstract

The Minnesota Department of Education (MDE) contracted for an outside alignment study of its Minnesota Comprehensive Assessment-II (MCA-II) for grades 3-8 and 11 using procedures based on the alignment model developed by Norman Webb. A panel of nine independent educators carried out three alignment tasks: Rating state benchmarks in mathematics for Cognitive Level A, B or C; rating core test items from MCA-II math tests for Cognitive Level A, B or C; and mapping test item hits for each benchmark. These ratings were variously applied to four alignment criteria: *Cognitive consistency, categorical concurrence, range-of-knowledge*, and *balance-of-representation*. Anecdotal feedback from alignment panels about the tests and standards was also reported.

The results show that the 2006 MCA-II were highly aligned for *categorical concurrence* and *range-of-knowledge*. Alignment for *cognitive consistency* had mixed results because the Panel rated most of the benchmarks as B- or C-level, leaving few matches for A-level test items. This was not considered a serious problem for the MCA-II tests because the tests require a substantial number of A-level items to measure basic skills; furthermore, a test geared just to the B- and C-level benchmarks would be impractical for the general population of students. Alignment for *balance-of-representation* was also mixed because many of the test items repeatedly matched a limited set of benchmarks. This was also not considered a serious problem because most of Minnesota's benchmarks were not intended for assessment by the MCA-II, but are left to local districts for classroom-based assessment. Three implications for follow-up actions were given.

An Alignment Study of the Minnesota Comprehensive Assessment-II with State Standards in Mathematics for Grades 3-8 and 11

The *No Child Left Behind Act* (NCLB) requires state education agencies to ensure their assessment systems are aligned with state academic standards. A state's alignment procedures are examined by the U. S. Department of Education through a Peer Review process for compliance with NCLB. For the purposes of this Peer Review, the Minnesota Department of Education (MDE) contracted with an independent specialist to conduct an alignment study of the core test items from the Minnesota Comprehensive Assessment-II (MCA-II) with state standards in mathematics for grades 3-8 and 11. A separate report describes a companion alignment of state standards in reading and literature with the MCA-II. These alignment studies were conducted during the summer months of 2006.

MDE's alignment procedures are based on the widely influential model developed by Norman Webb (1997, 1999) with some modifications. This approach has two avenues for alignment: The category of content covered by the state's content standards and assessments, and the complexity of knowledge required by these standards and assessments. Alignment for these purposes is operationally defined as an objective, independent process that determines the degree to which state standards and assessments are consistent for cognitive demand and academic content. A panel of independent experts, typically made up of master teachers, initially rates test items and academic standards for degree of cognitive demand, then maps concordance of content between each test item and the elements of the standards.

Webb contends that an alignment study for NCLB purposes "is not a simple yes or no analysis" (Webb, 2004a, p. 7). In order to have useful, formative data about the relationship between tests and standards, alignment must go beyond a superficial comparison of test items and academic content. Toward that end, Webb utilizes four alignment criteria (with modifications here to suit MDE's terminology). More detailed explanation of these calculations for criterion levels can be found at Webb (1999, 2004b):

Cognitive consistency compares coded ratings of cognitive complexity in each content standard and test item. Consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards. The criterion for consistency is met when at least 50% of test item hits are at or above the cognitive level specified in the corresponding content standard (levels A, B or C in Minnesota).

Categorical concurrence provides a very general indication whether both tests and standards incorporate the same content. It is judged by the number of test items for each standard, typically at least 6 test items per standard, in order to achieve an acceptable level of alignment. MDE has also used the criterion of 15% of the item pool when the standard of 6 test items is impractical (e.g., when there are numerous state achievement standards and a relatively small item pool). Early alignment studies under NCLB sometimes overly relied upon *categorical concurrence* data in lieu of more comprehensive criteria, such as those which follow.

Range-of-Knowledge is used to examine whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items. The criterion for this correspondence between span of knowledge for a standard is based on the number of benchmarks within the standard and matching test items. *Range-of-Knowledge* is met if 50% or more of the benchmarks for a standard have at least one related test item.

Balance-of-Representation is a proportional index that represents the distribution of content domains between content standards and assessments. Using the formula below, the distribution of assessment items is computed by considering the difference in the proportion of benchmarks and the proportion of hits assigned to the benchmark:

$$\text{Balance-of-Representation Index} = 1 - (\sum \left| 1/B_{k=1} - I_k/H \right|)/2$$

Where B = Total number of benchmarks hit for the standard
I_k = Number of items corresponding to the benchmark
K = Benchmarks
H = Total number of items hit for the standard

An index value of 1.0 signifies perfect balance, in which the corresponding items are equally distributed among the benchmarks and each benchmark is measured with the same number of items. The acceptable level on this criterion is .70.

Method

The methodology for this alignment uses an independent panel of experts to examine MCA-II tests in mathematics and the corresponding state content standards for mathematics. For aligning cognitive demand, benchmarks and test items are matched with a Likert-type scale based on Bloom's Taxonomy that represents a hierarchy of lower to higher order thinking skills. For aligning content, protocols were designed to map correspondence between state standards and test items. The alignment analyses use (or slightly modify, as explained below) Webb's recommended criteria (Webb, 1999).

Instruments

Bloom's Taxonomy Alignment Scale: Three Level Version. In previous NCLB alignments MDE used Webb's cognitive scale based on four levels of Depth-of-Knowledge, but in 2004 switched to an alignment scale based on Bloom's Taxonomy (Bloom, 1956). MDE found that a Bloom-based scale was more familiar and instructionally relevant to the independent panels of educators who make the alignment ratings (see MDE, 2004a, 2004b). A Bloom-based scale also proved beneficial at the front end of test development with outside vendors because the Cognitive Domain of Bloom's Taxonomy has been used for many years in developing curricula, instructional strategies, and assessments of student learning. An earlier alignment report (MDE, 2004a) describes the origin of MDE's Bloom-based scale, similar to one developed by

Florida's state education agency. Since Bloom's Taxonomy has several possible configurations for an alignment scale, a flexible title was chosen to depict the version as used here: Bloom's Taxonomy Alignment Scale: Three Level Version (BTAS-3). A BTAS version could potentially have as many as six levels, one for each of Bloom's cognitive descriptors, but that is impractical for a pencil-and-paper test primarily based on multiple choice test items. The viability of the BTAS for both reading and mathematics is addressed in earlier MDE alignment reports (2004a, 2004b).

For test development purposes, MDE condensed Bloom's six cognitive descriptors into three levels of cognitive demand: Cognitive Levels A, B and C (Figure 1). After trying various configurations and reviewing alignment efforts in other states that similarly used Bloom's Taxonomy, the three-level version in Figure 1 was strongly recommended via feedback from teacher panels for the following reasons:

1. State benchmarks and test items that primarily match Bloom's *Knowledge* and *Comprehension* categories are hierarchically distinct and should be separated into Cognitive Levels A and B, not combined (as was previously tried in MDE alignments).

2. It proved reasonable to combine the four highest categories in Bloom's Taxonomy—*Application*, *Analysis*, *Synthesis*, *Evaluation*—into a single Level C. An anticipated problem with overgeneralizing into Level C's four categories did not occur (see MDE, 2004a, p. 2). Due to time and other constraints of a statewide pencil-and-paper assessment, the most active skills at Level C will be *Application* (math) and *Analysis* (math and reading). Relatively few, if any, test items will primarily match *Synthesis* or *Evaluation* descriptors.

3. This scale version is useful for both reading and mathematics, thus simplifying alignment reporting for policymakers.

MCA-II. The MCA-IIs are the latest version of a series of criterion-referenced, or standards-based, tests that Minnesota schools have been administering since 2000. In accordance with test specifications prepared by MDE, private vendors were contracted to develop MCA-II tests that will provide information about how well students have learned the knowledge and skills set forth in academic standards passed by the Minnesota Legislature in 2003. This study examines the core test items in 2006 MCA-IIs in mathematics for grades 3-8 and 11. For these grades, the number of core test items ranges from 41 to 65.

Figure 1. Levels of cognitive demand for student learning in the BTAS-3.

Cognitive Demand	Matching Hierarchical Descriptors from Bloom's Taxonomy
Cognitive Level A represents the lowest level of complexity.	*Knowledge*: Remembering (recalling) of appropriate, previously learned information like terminology, specific facts, principles and generalizations. Test item cues include list, define, tell, describe, identify, label, collect, name, who, when, where.
Cognitive Level B requires an intermediate level of thinking.	*Comprehension*: Grasping (understanding) the meaning of informational materials. Test item cues include summarize, describe, interpret, contrast, discuss, estimate, distinguish.
Cognitive Level C is made up of Bloom's highest categories of cognitive complexity.	*Application*: The use of previously learned information in new and concrete situations to solve problems that have single or best answers. Test item cues include apply, demonstrate, calculate, complete, discover, solve, experiment, relate *Analysis*: Breaking down informational materials into their component parts, examining such information to develop divergent conclusions by identifying motives or causes, making inferences, finding evidence to support generalizations. Test item cues include analyze, separate, explain, connect, compare, infer, classify, order. *Synthesis*: Creatively or divergently applying prior knowledge and skills to produce a new or original whole. Test item cues include combine, integrate, modify, rearrange, substitute, design, formulate, generalize. *Evaluation*: Judging the value of material based on personal values/opinions, resulting in an end product, with a given purpose, without real right or wrong answers. Test item cues include assess, decide, rank, grade, test, measure, select, conclude, compare, explain.
A fourth point on the rating scale is "Not Ratable," should the raters determine that a benchmark does not sufficiently align at any level with Bloom's cognitive categories.	None.

State Benchmarks for Mathematics. Minnesota's academic content standards have the format displayed in Figure 2, where the example of "Number Sense, Computation and Operations" is one of five broad standards, or strands, for mathematics.[1] Underlying each sub-strand is an array of benchmarks ranging in number from 0-12; in several instances, benchmarks are duplicated or worded very closely among grade levels. Minnesota standards may be viewed on-line at http://education.state.mn.us.

[1] Minnesota's five broad expectations for Mathematics are Mathematical Reasoning; Number Sense, Computation and Operations; Patterns, Functions and Algebra; Data Analysis, Statistics and Probability; and Spatial Sense, Geometry and Measurement.

Figure 2. Sample format for Minnesota's statewide content standards.

	GRADE FOUR
Strand	***I. NUMBER SENSE, COMPUTATION AND OPERATIONS***
Sub-strand	**A. Number Sense**
Expectations	Represent whole numbers in various ways to quantify information and solve real-world and mathematical problems. Understand the concept of decimals and common fractions.
Benchmarks	1. Read and write whole numbers to 100,000, in numerals and words. 2. Compare and order whole numbers. 3. Use fractions and decimals to solve problems representing parts of a whole, parts of a set and division of whole numbers by whole numbers in real-world and mathematical problems. 4. Use rounding and estimation with whole numbers to solve real-world and mathematical problems.

Rater's protocols. Protocols were developed for each of the alignment tasks to record the panel's ratings and note comments. Each grade level has a separate set of protocols. Due to their length and irregular size, copies of protocols are not appended to this report but may be available upon request.

Participants

A panel of nine persons served as raters over a four-day session. Candidates for the panel registered with MDE's Assessment Advisory Panel Database. Selections were based on expertise and experience in teaching math and familiarity with state assessments. All raters were separately employed as a teacher or administrator in a local school district. As outside persons not employed by MDE, raters were entitled to travel reimbursement and a small honorarium.

Design and Procedure

The alignment sessions started with an orientation covering definitions, an overview of the alignment process, and training with the rating scale on practice benchmarks and test items. A facilitated group process was used to complete three alignment tasks:

Alignment Task #1: Rate benchmarks for Cognitive Level A, B or C.
Alignment Task #2: Rate test items for Cognitive Level A, B or C.
Alignment Task #3: Map test item hits for each benchmark.

Webb's procedures allow for averaging ratings from individual panel members or using consensus, but MDE prefers the latter because experience showed that consensus reports have higher reliability. Panel members benefit from group discussion in reaching their judgments about test items and standards, and the professional discourse reinforces consistency and lessens the need to revisit ratings. The facilitator notes cases where consensus is not achieved and only majority vote prevails, but these exceptions tend to be infrequent.

Findings

Findings are reported in five sections, one for each of the four alignment criteria plus feedback from the Alignment Panels. Tables in these sections summarize the status of alignment criteria by grade and standard. Individual tables for each grade and standard are too voluminous to be included in this report and may be obtained by contacting MDE.

Cognitive Consistency

Alignment for *cognitive consistency* is examined by comparing the cognitive level assigned to benchmarks with that of their matching test items, i.e., "hits." Hit counts represent the number of test item matches with direct correspondence to the benchmark content of a standard. Webb's procedures allow raters to code one primary hit for a test item—if one is evident—and additional hits. Combining both primary and secondary hits between test items and content standards is important because it is commonplace for test items to be relevant to more than one benchmark (or standard)[2]. Combining primary and secondary hits sometimes produce hit counts that exceed the number of test items.

After benchmarks and test items were sorted into Level A, B or C, hits were tallied where test items matched each of the five standards. According to Webb's alignment model, at least 50% of matching test items are expected to rate at or above the same cognitive levels as their corresponding benchmarks to achieve cognitive consistency (Table 1). This criterion level was often unmet in this study because the Alignment panel rated most of the benchmarks at the B and C level, and the tests contain many items at A-level. The Panel gave very few benchmarks the A-level rating because many of them are "all encompassing" by consolidating A- and B-level skills, sometimes A-, B- and C-level skills. Since the Panel had to select one level per benchmark, it chose to align to the highest level represented—Hence, there were few A's and many B's and C's to match with test items.

[2] In fact, the strand for Mathematical Reasoning was not intended to stand alone but to complement the content of other strands that appears in test items.

This finding is not perceived as a serious alignment problem in this case because pencil-and-paper tests *should* have several A-level items, especially at the lower grades. Furthermore, standardized tests of this type typically do not have sufficient time or resources available for students to work out highly complex test items requiring Synthesis and Evaluation skills. Local classroom-based assessment is routinely used for assessing these highest categories in Bloom's Taxonomy. If the alignment criterion for *cognitive consistency* had been routinely met in this study, the tests would be extraordinarily difficult for the general population because the state chose to have predominantly complex, highly challenging benchmarks (as rated by the Panel).

Table 1. Summary of cognitive consistency for mathematics, grades 3-8 and 11

Standard	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Mathematical Reasoning	Yes	No	Yes	No	No	No	No
Number Sense, Computation, and Operations	No	Yes	Yes	Yes	No	No	Yes
Patterns, Functions, and Algebra	No	No	No	Yes	No	No	No
Data Analysis, Statistics, and Probability	Yes	No	Yes	No	No	No	No
Spatial Sense, Geometry, and Measurement	Yes	Yes	No	No	No	Yes	Yes

Categorical Concurrence

Categorical concurrence is a general indicator of content matching that calls for at least six matches between test items and academic standards. Table 2 shows that this criterion was virtually met for all standards at all grade levels, with only one marginal finding for Mathematical Reasoning at grade 3 (which had five matches). It should also be noted that *categorical concurrence* is aligned by counting hits at the strand level, while other alignment criteria tally hits at the benchmark level.

Table 2. Summary of categorical concurrence for mathematics, grades 3-8 and 11

Standard	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Mathematical Reasoning	Marginal	Yes	Yes	Yes	Yes	Yes	Yes
Number Sense, Computation, and Operations	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patterns, Functions, and Algebra	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data Analysis, Statistics, and Probability	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spatial Sense, Geometry, and Measurement	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Range-of-Knowledge

Range-of-Knowledge provides more comprehensiveness to the alignment analysis than categorical concurrence, since the latter could be met by having six test items match only one or two out of several benchmarks. *Range-of-Knowledge* indicates the span of content covered by a test by requiring 50% or more of a standard's benchmarks to have at least one related test item. Table 3 shows that this alignment criterion was widely met for all grade levels and standards, only missing at Grade 8 for Mathematical Reasoning and Grade 11 for Number Sense, Computation and Operations.

Table 3. Summary of range-of-knowledge for mathematics, grades 3-8 and 11

Standard	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Mathematical Reasoning	Yes	Yes	Yes	Yes	Yes	No	Yes
Number Sense, Computation, and Operations	Yes	Yes	Yes	Yes	Yes	Yes	No
Patterns, Functions, and Algebra	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data Analysis, Statistics, and Probability	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spatial Sense, Geometry, and Measurement	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Balance-of-Representation

Balance-of-Representation adds comprehensiveness to an alignment analysis because it complements *range-of-knowledge*. While *range-of-knowledge* reflects the span of content covered by test items, one or two benchmarks sometimes net most of the hits which serves to limit the breadth of alignment. Hence, the *balance-of-representation* index adds to the alignment by indicating the spread of test items across the array of benchmarks. Table 4 shows that the criterion index of .70 was met or marginally met in just over half the instances. The Grade 4 test had balance among the benchmarks for all the standards, while the Grade 11 test did not meet the alignment criterion for any standards. These mixed results may be somewhat misleading because there are several reasons for these patterns.

The primary reason the test items matched a limited set of benchmarks is that many benchmarks were not intended for the MCA-II; most of Minnesota's benchmarks were intended for classroom-based assessment. Therefore, many benchmarks were ignored while certain ones were repeatedly tallied by the alignment panel, giving the impression of limited balance among the entire pool of benchmarks. Another artifact creating the impression of limited balance is the fact that some sub-strands have zero, one or two benchmarks, a condition which complicates the use of Webb's formula for this criterion. Also, there were relatively few hits on Mathematical Reasoning benchmarks because there was overlapping language in the Algebra standard that was more encompassing.

Table 4. Summary of balance-of-representation for mathematics, grades 3-8 and 11

Standard	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Mathematical Reasoning	No	Yes	No	No	No	No	No
Number Sense, Computation, and Operations	No	Yes	No	No	Marginal	No	No
Patterns, Functions, and Algebra	No	Yes	Yes	Yes	No	Yes	No
Data Analysis, Statistics, and Probability	Yes	Yes	Yes	Marginal	Yes	Marginal	No
Spatial Sense, Geometry, and Measurement	Yes	Yes	Yes	Yes	Yes	Marginal	No

Panel Feedback

Equally valuable to the alignment ratings was feedback from Alignment Panel about the tests and standards. Listed below are recommendations or concerns brought up, some of which echo feedback from previous alignment studies:

1. Grade 11 benchmarks are unclear as to underlying basic skills required. If this was clearer, it would help teachers in lower grades adjust their instruction.

2. The Panel had adverse comments or recommendations for several test items:
   - Grade 11, item 11 is confusing because the use of a commas makes it appear that a digit is missing from a number
   - Grade 6, item 15 has unclear direction in statement, Each operation must be used as least once"
   - Grade 6, item 23 is problematic by the double use of the word "strip"
   - Grade 5, item 13 uses the word "ordered", which is a math term with a different meaning
   - Grade 3, item 9 has directions that are too wordy plus words "list" and "show" mean different things so use one verb

3. Benchmarks for Data & Statistics in Grade 3 should not be limited to the one or two kinds of figures (e.g., circle).

4. Benchmark #2 for Spatial Sense in Grade 8 should cover two dimensional objects as well as three dimensions.

5. Benchmark #2 for Computation & Operation in Grade 6 should use the word "factor" instead of "divisor"

Implications and Discussion

The methodology of using consensus ratings by a panel of experts was comparable to previous alignment studies. Professional discourse was successful at resolving differences among the panel members, resulting in consensus agreement on all elements of the three alignment tasks. Minnesota's BTAS-3 continued to be useful for alignment purposes in rating the depth of cognitive demand for both state content standards and test items. The A, B, C ratings for the mathematics benchmarks appeared successfully used as a baseline for initially developing test specifications, and again in this study to compare with test items.

The findings for *cognitive consistency* were mixed across the grade levels. This occurred primarily because the Alignment Panel rated very few benchmarks at Level A. Due to the tendency of the benchmarks to incorporate multi-level skills, the Panel rated benchmarks according to the highest inclusive skill. Therefore, virtually all the Level A test items were unmatched to benchmarks, giving the appearance of inconsistency. However, Level A elements of the benchmarks were actually covered by numerous test items but it is not evident via this alignment methodology. To illustrate, consider the following Grade 5 benchmark for Patterns and Functions:

> *Identify patterns in numbers, shapes, tables, and graphs and explain how to extend those patterns.*

This benchmark has both Level A and C components ("identify patterns" and "extend patterns"), but it got an overall rating of C because the Alignment Panel chose to rate multiplicitous benchmarks at the highest skill level included. Therefore, Level A test items that required only basic identification of patterns do not get credited for consistency with this Level C benchmark.

Implication #1. Minnesota's benchmarks could be revised to more clearly delineate A-, B- and C-level cognitive skills, instead of consolidating hierarchical skills across a single benchmark.

Categorical concurrence and *range-of-knowledge* were uniformly demonstrated for all grade levels and standards. There are no implications pursuant to these alignment criteria.

The findings show inconsistent to poor *balance-of-representation* because many benchmarks were not hit by test items. The extent to which this is a problem should be investigated further, since many of the benchmarks are unsuited to pencil-and-paper standardized testing. Therefore, this alignment criterion should be limited to examining balance among those benchmarks intended for coverage by MCA-II tests.

Implication #2. A follow-up study could sort out the "essential benchmarks" intended for statewide assessment by the MCA-II tests and recalculate the *balance-of-representation* indices for each grade level.

In addition to the implications regarding the four alignment criteria, the anecdotal feedback from the alignment panels suggests that the standards need some revisions. In addition to correcting typographical errors, the state's format bears reconsideration so that learning and instructional expectations are clearer for near-identical benchmarks at multiple grade levels.

Implication #3. The alignment panel's recommendations may provide constructive input to the group undertaking the revision of Minnesota's standards in mathematics.

References

Bloom, B.S. (Ed.) (1956) Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York ; Toronto: Longmans, Green.

Minnesota Department of Education. (2004a). An alignment of Minnesota's benchmarks in reading & literature for grades 3, 5, 6, 8 and 10: June 2004.

Minnesota Department of Education. (2004b). An alignment of Minnesota's benchmarks in mathematics for grades 3, 5, 6, 8 and 11: July 2004.

Webb, N. L (2004a). Results of the (Delaware) Alignment Study for the Content Standards in Mathematics Grades 3,5,8, and 10. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L (2004b). Findings of an Alignment Study in DSTP English LanguageArts and Mathematics for Grades 3, 5, 8, and 10; and Science for Grades 4, 6, 8, and 11. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessment in four states* (Monograph No. 18). Madison: University of Wisconsin, Council of Chief State School Officers and National Institute for Science Education Research.

Webb, N. L. (1997). Research Monograph #6: Criteria for alignment expectations and assessments in mathematics and science education. Washington, D.C.: Council of Chief State School Officers.