# An Alignment Study of the Minnesota Comprehensive Assessment-II with State Standards in Reading and Literature for Grades 3-8 and 10

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

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 10 using procedures based on the alignment model developed by Norman Webb. A panel of eight independent educators carried out three alignment tasks: Rating state benchmarks in reading and literature for Cognitive Level A, B or C; rating core test items from MCA-II reading 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 tests were highly aligned for *cognitive consistency*, *categorical concurrence* and *range-of-knowledge.* Alignment for *balance-of-representation* was dropped from the study because the Alignment Panel recorded a very high number of duplicate benchmark to test item hits, which rendered the formula unusable. The tendency toward high duplicate hit counts was due to the Panel's strongly held view that Minnesota has a series of repetitive benchmarks that are so broadly inclusive they readily apply to almost any test item. Due to the positive alignment findings for the remaining three criteria, there are no implications or recommendations other than options for reexamining *balance-of-representation.*

An Alignment Study of the Minnesota Comprehensive Assessment-II with State Standards in Reading and Literature for Grades 3-8 and 10

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 reading and literature for grades 3-8 and 10. A separate report describes a companion alignment of state standards in mathematics 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 reading and the corresponding state content standards for reading and literature. 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 for the Panel 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 reading for grades 3-8 and 10. For these grades, the number of core test items ranges from 40 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<br><br>*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.<br><br>*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.<br><br>*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 Reading and Literature.* Minnesota's academic content standards have the format displayed in Figure 2, where the example of "Vocabulary Expansion" is the second of four broad expectations, or sub-strands, for reading and literature.[1] Underlying each sub-strand is an array of benchmarks ranging in number from

[1] Minnesota's four broad standards for Reading and Literature are Word Recognition, Analysis and Fluency; Vocabulary Expansion; Comprehension; and Literature. The standard for Word Recognition, Analysis and Fluency is not included in this alignment study because test specifications for MCA-IIs did not call for matching test items. With its emphasis on oral reading skills, this particular standard is intended for assessment by teachers at the classroom level and not through a statewide pencil-and-paper test. In addition, the state leaves the content of this standard entirely to local discretion for grades eight and higher. Therefore, it was moot to apply alignment criteria to Word Recognition, Analysis and Fluency.

1-13; in several instances, benchmarks are duplicated or worded very closely among grade levels. For this study each benchmark was compared, or aligned, with the BTAS-3 to represent the level of cognitive skills required by students to meet that learning expectation. Minnesota standards may be viewed on-line at http://education.state.mn.us.

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

| ***GRADE 4*** | |
|---|---|
| *Strand* | **I. READING AND LITERATURE**<br>The student will read and understand grade-appropriate English language text. |
| *Sub-strand* | **B. Vocabulary Expansion** |
| *Standard* | Standard: The student will use a variety of strategies to expand reading, listening and speaking vocabularies. |
| *Benchmarks* | The student will:<br>1. Acquire, understand and use new vocabulary through explicit instruction and independent reading.<br>2. Identify and understand root words, derivations, antonyms, synonyms, idioms, homonyms and multiple-meaning words to determine word meanings and to comprehend texts.<br>3. Use dictionaries and glossaries to understand the meaning of new words.<br>4. Use context and word structure to determine word meanings.<br>5. Use knowledge of prefixes and suffixes to determine the meaning of unknown words. |

*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 eight 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 reading 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 secondary 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). For example, a test item could align with benchmarks from the comprehension and vocabulary standards.[2] Combining primary and secondary hits often produce hit counts on reading tests 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 for all grade levels and standards in this study except for Grade 10 Vocabulary.

[2] This sometimes happens with Minnesota standards for two reasons. Some benchmarks are so broadly worded they readily net a lot of multiple its. Also, some content is shifted around among the standards, e.g., skills with similes and metaphors may variously appear in either the comprehension, literature or vocabulary benchmarks at different grade levels.

Table 1. Summary of cognitive consistency for reading and literature, grades 3-8 and 10

| Standard | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | Grade 10 |
|---|---|---|---|---|---|---|---|
| Vocabulary Expansion | Yes | Yes | Yes | Yes | Yes | Yes | No |
| Comprehension | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Literature | Yes | Yes | Yes | Yes | Yes | Yes | No |

*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 met for all standards at all grade levels. 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 reading and literature, grades 3-8 and 10

| Standard | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | Grade 10 |
|---|---|---|---|---|---|---|---|
| Vocabulary Expansion | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Comprehension | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Literature | Yes | Yes | Yes | Yes | Yes | Yes | No |

*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 met for nearly all grade levels and standards. It is noted that the term "marginal" is used in Table 3 when the hits were one away from meeting the 50% criterion. This was done because there are several instances where there is an odd number of benchmarks (7, 9, etc.), and a 50% criterion literally cannot be met.

Table 3. Summary of range-of-knowledge for reading and literature, grades 3-8 and 10

| Standard | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | Grade 10 |
|---|---|---|---|---|---|---|---|
| Vocabulary Expansion | Yes | Yes | No | Yes | Yes | Marginal | Yes |
| Comprehension | Yes | Yes | Yes | Yes | Marginal | Marginal | Yes |
| Literature | Marginal | Marginal | Yes | Yes | No | No | Yes |

*Balance-of-Representation*

The *balance-of-representation* criterion turned out to be unusable in this study because of the very high number of multiple hits recorded by the Alignment Panel. Webb's formula for *balance-of-representation* anticipates one primary hit between a test item and a benchmark, allowing for the possibility of up to two more secondary hits. The Panel in this study determined that the broad phrasing of certain benchmarks in Comprehension and Literature created so many frequent, duplicate hits that the sheer number confounded Webb's formula. This effect can be seen in the formula below:

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

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

In this formula, a high number of duplicate hits (B) will overwhelm the number of test items (H), creating a spurious proportional index that is unusable. In brief, the problem is that the initial test specifications identify one particular benchmark as the target for each test item, but the Panel may have identified as many as five benchmarks as major hits. The Panel felt this was unavoidable because Minnesota has certain benchmarks so broadly worded that they readily apply to almost any test item.[3] The following Grade 5 benchmarks illustrate this view:

> *Generate and answer literal, inferential, interpretive and evaluative questions to demonstrate understanding about what is read.* (Comprehension, #7)
>
> *Respond to literature using ideas and details from the text to support reactions and make literary connections.* (Literature, #8).

[3] Previous Alignment Panels have expressed the same view and showed the same tendency toward duplicate test item-to-benchmarks hits, though to a lesser degree that did not overwhelm the *balance-of-representation* formula.

Not only were benchmarks like these perceived as broadly inclusive, they are repeated at most grade levels and that compounds the tendency toward very high hit counts for most test items. Therefore, the *balance-of-representation* indices were dropped from this study in lieu of other options such as convening a second panel to redo the ratings or modifying the formula to accommodate high hit counts.[4]

*Panel Feedback*

Feedback from alignment panels about state tests and standards is often as valuable as the alignment ratings. This Panel had numerous recommendations and comments for MDE to consider in revising state standards and the MCA-II tests. MDE will be provided separately with a complete list, but as it is rather lengthy it will be summarized below:

1. The Panel had comments or recommendations for several test items such as the following:
    - Not enough context to answer question
    - No clear answer in the passage
    - Item has two possible answers
    - One item cues the answer for another item
    - Grade 3, item 9 has directions that are too wordy plus words "list" and "show" mean different things so use one verb

2. The Panel had comments or recommendations for several benchmarks such as the following:
    - Ambiguously worded and difficult to measure
    - Need comprehension benchmarks for main idea, cause and effect, supporting details, sequencing, summarize and paraphrase
    - The intent of slight wording changes on benchmarks from grade to grade is confusing (word "analyze" at lower grade connotes higher level cognitive skill than "determine")
    - Clarify what forms are included in the term "literature"
    - Benchmarks 8 and 13 are basically the same (Grade 6 Comprehension)
    - From grade to grade, skills for using similes and metaphors seem arbitrarily applied to Comprehension, Vocabulary, or Literature

## 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 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 reading and literature benchmarks appeared successfully used as a baseline for initially developing test specifications, and again in this study to compare with test items. Due to the positive alignment findings for

[4] This may be done by changing the value for H to $\sum I k$.

three of the four criteria (*categorical concurrence, cognitive consistency, range-of-knowledge*), there are no implications or recommendations other than options for reexamining *balance-of-representation.*

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.