Alignment of the MCAS Assessments To State Content Standards: An Example\textsuperscript{1, 2}

Stephen J. Jirka and Ronald K. Hambleton
University of Massachusetts Amherst

September 30, 2005

\textsuperscript{1} Center for Educational Assessment MCAS Validity Report No. 12. (CEA-576). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

\textsuperscript{2} This report was completed under a 2004-2005 contract between the Center for Educational Assessment at the University of Massachusetts and Measured Progress, Dover, New Hampshire.
Alignment of the MCAS Assessments To State Content Standards: An Example
Stephen J. Jirka and Ronald K. Hambleton
University of Massachusetts at Amherst

Introduction and Overview

To comply with the federal No Child Left Behind (NCLB) law, states must provide evidence that they are following numerous requirements of the law. One section of the law requires the alignment of academic content standards, academic achievement standards (or performance standards), and the assessments themselves. This is important, since the assessment and accountability systems will be a valid measure only to the extent to which they are in accord with the standards that have been set up by the state. To help states, like Massachusetts, meet the requirements of NCLB, the US Department of Education produced a guide called Standards and Assessments: Peer Review Guidance (USDOE, 2004). This guide gives examples of the types of evidence states might provide to show that they are in compliance with the federal requirements. This brief report will provide some background information on (1) why it is necessary to have an independent alignment study, (2) what method might be used, (3) a sample protocol for an alignment study, and (4) what type of data such a study might provide. The goal of this short report is to offer a methodology that extends the work of Hambleton and Zhao (2004) with the grade 10 English Language Arts and Mathematics Assessments and is consistent with federal guidelines.
As the Peer Review Guide reminds us, “demonstrating that an assessment system is aligned with a State's standards requires more than simply determining whether all the items on the assessment can be matched to one or more standards or whether each of the academic content standards can be matched to one or more items in the assessments.” (USDOE, 2004). The Guide goes on to list the components or features that states like Massachusetts should consider when ensuring its curriculum frameworks, learning standards, and assessments are aligned. The assessment must

- cover the full range of content specified in the curriculum frameworks and related learning standards, meaning that all of the learning standards are represented legitimately in the assessments
- measure both the content (what students know) and the process (what students can do) aspects of the curriculum frameworks
- reflect the same degree and pattern of emphasis apparent in the curriculum frameworks and learning standards (e.g., if the learning standards place a lot of emphasis on writing, then so should the assessments)
- reflect the full range of cognitive complexity and level of difficulty of the concepts and processes described, and depth represented, in the curriculum frameworks and related learning standards, meaning that the assessments are as demanding as the learning standards
- yield results that represent all performance categories used by the state in classifying students and reporting scores (in Massachusetts, there are four performance levels: Failing or Warning, Needs Improvement, Proficient, and Advanced)
The *Peer Review Guide* emphasizes that each state must present evidence that its assessment system is aligned to its curriculum frameworks and learning standards at many different time points in the development of a complete assessment system. As the assessment system is developed, for example, some alignment evidence is generated in the test development process, and documentation of the steps taken to ensure that items were drafted to reflect the full range of the state standards is appropriate verification of efforts to attain alignment.

In this report, we will concentrate on a procedure that could be implemented in 2005-2006 on the last five years of the assessments in English Language Arts and Mathematics—2001 to 2005. Namely, we discuss using UMass and/or the Massachusetts Department of Education’s curriculum and test development committees to show the alignment of the assessments with the state curriculum frameworks and learning standards. If necessary, documentation could be produced showing that the changes that might have been needed to improve the alignment have been completed. The *Guide* mentions the necessity of “reports of independent alignment studies and evidence of a systematic process for addressing any gaps or weaknesses identified in these studies” (USDOE, 2004).

In recent years, several methods of evaluating alignment between content standards and assessments have been developed. A summary and comparison of alignment models can be found in a report by Li and Sireci (2004) completed in the fall for the 2004-2005 UMass-Measured Progress contract. The *Guide* (USDOE, 2004) also lists what should be documented and described when ensuring the alignment between standards and the state assessment system:
1. The relationships between the structure of the standards and the structure of the assessments

2. The rationale for the overall alignment strategy, including a rationale for any strands or learning standards in the curricula either not assessed or not reported as part of the state assessment

3. The manner in which each strand and standard is assessed, whether at the state, district, school, or classroom level

4. The manner in which alternate assessments based on alternate performance standards are linked to the state content standards

5. The type of information the State collects pertaining to each standard, and how the State monitors the quality of the assessment data collected at the local level, for all assessments that are part of the statewide system

**Models for Alignment Study**

As pointed out by Li and Sireci (2004), many different models are available to examine the alignment between the state standards and assessments, and those models are more fully described in that article. The Webb and the Achieve models are the two most widely used. Here we note that the Webb model certainly has been widely used (10 states at least) to produce data states can use to provide evidence needed for NCLB. However, the model is complicated, and only the subcategories of the Content Focus criterion were applied in Webb’s own empirical alignment studies. Another attractive model that can fulfill the evidentiary requirements of NCLB is the model employed by Achieve, Inc., and a sample protocol is provided next that shows how such an alignment study could be implemented by UMass/Department of
Education. Since 1998, this model has been developed, tested, and used in over a dozen states (Achieve, 2004), and is the one we prefer.

Sample Protocol

The Peer Review Guide mentions that states, like Massachusetts, must not only show how the assessment matches the relevant state curriculum framework as defined by strands and learning standards, but also what the perceived strengths and weaknesses are. The Achieve model enables the state to gain greater insight into the assessment’s strengths and weaknesses as they are analyzed using a set protocol for each grade and content area.

In order to preview and clarify some of the terms to be used, the five criteria that are central in determining the degree of alignment between content standards and an assessment are given below (Rothman et al., 2002), and a sample item rating sheet is shown at the end of this report for grade 8 mathematics.

1. Confirmation of the state's test blueprint: The blueprint is the state’s official statement of alignment and shows the intended match between each test item and a learning standard. We have this information for five tests we might work with during the 2005–2006 year (ELA, grades 4 and 7; Mathematics, grades, 4, 6, and 8). This first step is an important check on the test developer's judgment because test blueprints are typically used by states as the basis for their score reports. Essentially, the question asked is, "Is the assignment of an item to a learning standard justified?" Then one makes a "yes" or "no" decision.

2. Content centrality: This criterion further analyzes the match between the content of each test item and the content of the related learning standard by examining the
degree or quality of the match. Both the specificity of the learning standard and the extent to which the content to be assessed is evident from a reading of the item are considered. Reviewers assign each test item to one of four categories based on the perceived degree of alignment. These categories are described later in this report.

3. **Performance centrality**: This criterion focuses on the degree of the match between the type of performance (cognitive demand) presented by each test item and the type of performance described by the related standard. Each item makes a certain type of cognitive demand on a student (e.g., the item requires a certain performance such as "select," "identify," "compare," or "analyze"). The key issue is whether there is consistency between the type(s) of performance demanded by the item and the type(s) of performance described in the learning standards. Reviewers assign each item to one of four categories based on the degree of alignment. Again, these are described later in this report.

4. **Challenge.** This criterion is addressed at both the item level and for the set of items that maps to a learning standard. Its purpose is to determine whether doing well on these items requires students to master challenging subject matter. At the item level, reviewers consider two factors related to challenge: source of challenge and level of cognitive demand.

   *Source of challenge* attempts to uncover whether the individual test items in a set are fair. It seeks to establish whether items are difficult because of the knowledge and skills they target, or for other reasons not related to the subject matter, such as relying unfairly on students' background knowledge. Reviewers rate each item as having an appropriate or inappropriate source of challenge, and any item judged to have an
inappropriate source of challenge is not examined when item sets are evaluated for level of challenge.

*Level of cognitive demand* focuses on the type and level of thinking and reasoning required by the student on a particular test item. A Level 1 (recall) item requires the recall of information such as a fact, definition, term, or simple procedure. A Level 2 (skill/concept) item calls for the engagement of some mental process beyond a rote response, with students required to make some decisions as to how to approach a problem or activity. Level 3 (strategic thinking) items require students to reason, plan, or use evidence. Level 4 (extended thinking) items require complex reasoning, planning, developing, and thinking, most likely over an extended period of time. Here, there are a number of other systems that might be used for describing cognitive demand including Bloom’s taxonomy (six levels), or the systems used by NAEP (e.g., in mathematics, NAEP has used three levels to describe cognitive levels--conceptual understanding, procedural knowledge, and problem-solving).

Level of challenge applies to the set of items that map to a standard. This criterion compares the emphasis of performance required by a set of items to the emphasis of performance described by the related learning standard. In addition to evaluating alignment, reviewers also judge whether the set of test items has a span of difficulty appropriate for students at a given grade level based on the learning standards, the assessments and the supporting materials. Reviewers are asked to write a succinct summary of the level of challenge of each item set. It is requirements such as this one that lead to the conclusion that assistance from the Department of Education’s curriculum and test development committees will be needed.
5. **Balance and range**: These criteria are applied to the set of items that map to a learning standard. **Balance** compares the emphasis of content and performance supplied by an item set to the emphasis of content and performance described by the learning standards. In addition to evaluating alignment, reviewers also judge whether the set of items emphasizes the more important content at the grade level. Again, we think the curriculum and test development committees will be in a stronger position than UMass to address this concern. Reviewers write a succinct summary of the balance of each item set. **Range** is a measure of coverage or breadth (the numerical proportion of all content addressed) in a learning standard.

Because no single assessment can measure the full range of knowledge and skills required by the state learning standards, the Achieve model provides both qualitative and quantitative descriptive information about the choices states and test developers have made by looking at the balance and range of the assessment. This criterion was addressed well in the Hambleton-Zhao study carried out last year. The process itself brings out different factors that contribute to the alignment between assessments and content standards, thereby providing the types of information suggested in the *Peer Review Guide* (USDOE, 2004).

The detailed scoring criteria are described next for content centrality, performance centrality, and source of challenge. First, the reviewers score each item for **content centrality** as follows:

- 2 clearly consistent
- la not specific enough (learning standard or objective is too general to be assured of an item's clear consistency)
Each item for *performance centrality* is scored as follows:

2   clearly consistent  

1a  somewhat consistent (objective is too general to be assured of an item's clear consistency)  

1b  somewhat consistent (learning standard or objective uses more than one verb, but the item matches the less cognitively demanding of the verbs)  

0   inconsistent  

Finally, reviewers rate each test item for source of challenge as follows:

1   appropriate source(s) of challenge  

0   inappropriate source(s) of challenge  

Additionally, reviewers provide an explanation for items receiving a "0" for source of challenge and do not include such test items in evaluation level of challenge.

When evaluating the level of cognitive demand, the reviewer asks the following central question: What is the level of cognitive demand of the test item? Judgments about level of cognitive demand focus on the type and level of thinking and reasoning required by students on individual items. To be more specific, four levels of cognitive demand, ranging from the simple recall of information to complex reasoning and thinking are employed in the Achieve system, and reviewers rate each item for level of cognitive demand. Examples of the levels in mathematics follow:
**Level 1** (recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. In mathematics, a one-step, well-defined, and straight algorithmic procedure is considered to be at this level.

**Level 2** (skill/concept) includes the engagement of some mental processing beyond a rote response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomena and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret" could be classified at different levels depending on the object of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is a Level 2 activity. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how well information from the graph can be aggregated is a Level 3 activity.

**Level 3** (strategic thinking) requires reasoning, planning, using evidence and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3 activity. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex
and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be classified as Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; and using concepts to solve problems.

**Level 4** (extending thinking) requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4 activity. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives to how the task should be solved/completed, in order to be at this highest level. Level 4 activities include designing and conducting experiments, making connections between a finding and related concepts and phenomena, combining and synthesizing ideas into new concepts, and critiquing experimental designs.

If we adopt these four cognitive levels for our work, then we would need to research the literature for explanations of the four levels in reading and writing.
Derived Summary Data

Summary tables such as the following can be constructed from data derived from the alignment study and submitted as evidence to the federal government of the alignment of the MCAS assessments.

Number of Eligible Content Standards Assessed and Not Assessed in Grades 4, 6, 8, and 10 Mathematics

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th>Grade 6</th>
<th>Grade 8</th>
<th>Grade 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number/Percent Assessed</td>
<td>14/48%</td>
<td>13/43%</td>
<td>13/46%</td>
<td>14/45%</td>
</tr>
<tr>
<td>Number/Percent Not Assessed</td>
<td>15/52%</td>
<td>17/57%</td>
<td>15/54%</td>
<td>17/55%</td>
</tr>
<tr>
<td>Total</td>
<td>29/100%</td>
<td>30/100%</td>
<td>28/100%</td>
<td>31/100%</td>
</tr>
</tbody>
</table>

Summary Percentages for Grades 4, 6, 8, and 10 Mathematics

<table>
<thead>
<tr>
<th>Grade</th>
<th># of</th>
<th>Content Centrality</th>
<th>Performance Centrality</th>
<th>Source of Challenge</th>
<th>Level of Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>la</td>
<td>lb</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>82%</td>
<td>1%</td>
<td>9%</td>
<td>7%</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>97%</td>
<td>0%</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>8</td>
<td>60</td>
<td>89%</td>
<td>0%</td>
<td>11%</td>
<td>0%</td>
</tr>
<tr>
<td>10</td>
<td>60</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>2</td>
</tr>
</tbody>
</table>

Summary

Evidence of alignment between the state standards and assessments is required under NCLB, and the US Department of Education’s Standards and Assessments: Peer Review Guidance (USDOE, 2004) gives examples of the types of evidence states might provide to show that they are in compliance with the law. This brief report provided some background information on why it is necessary to have an independent alignment study, what method might be used, a sample protocol for an alignment study, and what type of data such a study might provide.
We are ready to begin with MCAS assessments for grades 4 and 7 ELA and grades 4, 6, and 8 Mathematics. We have the tests printed and ready to be reviewed. We can also repeat the analyses we carried out on the grade 10 tests last year. Using the state’s own item-learning standard matches, we can address the extent to which (1) the assessments are in line with the test specifications and (2) the curriculum frameworks that are being assessed.

The big question though is whether the UMass team should direct the work being described with the Massachusetts curriculum and test development committees, or complete the work ourselves, and have the Massachusetts curriculum and test development committees review our work and respond to our omissions. There is also the question of how much the Achieve methodology can be modified and still have the state in compliance with federal government expectations.
References


Version: September 30, 2005