Arizona Mathematics Standard and Assessments Alignment

Jerome V. D'Agostino

July 2010

Arizona Mathematics Standard and Assessments Alignment

Executive Summary

A vital process in the validation of standards-based assessments is the examination of alignment between the tests and academic standards. In March, 2010, 23 subject matter experts (SMEs) reviewed the congruence between the Arizona Academic Standard in mathematics and the 2010 Arizona Instrument to Measure Standards (AIMS) mathematics assessments for grades three through eight and tenth. The Web Alignment Tool (WAT) was used to examine alignment in five dimensions: (1) Categorical Concurrence, (2) Depth of Knowledge consistency, (3) Range of Knowledge correspondence, (4) Balance of Representation; and (5) Source of Challenge. A committee of six SMEs judged the exams in grades three and four, six SMEs rated the tests in grades five and six, five SMES reviewed the grade seven and eight exams, and six SMEs scored the tenth grade test. The tests consisted of between 66 and 85 operational multiple-choice items.

Overall, the alignment results indicated rather strong concurrence between the state standards and assessments in mathematics, with excellent scores for Categorical Concurrence, Range of Knowledge, and Balance of Representation. There were some weaknesses indicated for Depth of Knowledge, revealing that SMEs at times did not rate the cognitive demand level of test items to be comparable to the objectives they matched with items.

Across the seven test levels and five strands of the state academic standard, AIMS received 34 of 35 "Yes" values for Categorical Concurrence. The only "No" conclusion was for the tenth grade test in Number Sense. To earn a "Yes" WAT score, a strand must have at least six items matched to it. The tenth grade test was designed with only five Number Sense items, which likely was the reason for the lack of Categorical Concurrence in that area. Follow up analyses revealed that judges matched three of the five items with Number Sense objectives.

In 22 of 35 cases, Depth of Knowledge was considered acceptable, with particular strengths in Number Sense and Geometry & Measurement. Five of the seven exams received "Yes" ratings for Patterns, Algebra, & Functions. Data Analysis & Probability and Math Structure & Logic received one and two "Yes" values, respectively. For the latter strands, SMEs did not judge the items to require the same cognitive demand levels as the objectives. Range of Knowledge had "Yes" values in 30 of 35 cases, with some weaknesses in Math Structure & Logic. Balance of Representation was deemed acceptable in all but one case.

In the WAT procedure, SMEs match items with objectives without reviewing test specifications, and Categorical Concurrence is not scored in consideration of the item-objective linkages designed by the test developer. Consequently, a test can receive "Yes" scores for a strand even if SMEs did not match items actually linked to the strand according to test specifications. The WAT does not provide an evaluation of item-objective maps. To address this issue, follow up Categorical Concurrence methods were developed that evaluated the degree to which SMEs matched items to strands as specified by the test item maps, and that considered the number of linked items per strand to determine if Categorical Concurrence was met.

These alternative methods revealed that SMEs matched the majority of linked items to strands. In all but one case, Math Structure & Logic in grade 7, the SMEs matched the majority of items to strands as specified in test blueprint documents. Overall, the WAT evidence clearly indicated positive alignment between the 2010 AIMS mathematics exams and Arizona academic standard in mathematics.

Arizona Alternate Academic Standards and Assessments Alignment

A critical step in validating standards-based assessments is to examine the congruence or alignment between test items and the standards for which they were designed to measure. Without sufficient alignment, standards-based reform ultimately will fail because the connection between a student's test score and the teacher's efforts to center instruction around state standards will be tenuous. On March 4th and 5th, 2010, 23 subject matter experts (SMEs) reviewed the congruence between the Arizona Academic Standard in mathematics and the 2010 Arizona Instrument to Measure Standards (AIMS) mathematics assessments for grades three through eight and tenth. The alignment study was conducted by Jerome V. D'Agostino, Associate Professor of Quantitative Methods at The Ohio State University. This report documents the characteristics of the SMEs who evaluated the assessment items and learning objectives that comprise the state standards, the methods used to collect the data, the results of the analysis, and conclusions and recommendations.

Aligning Tests and Standards

The alignment of tests and standards begins in the test construction process. After academic standards are established, states typically develop test blueprints that specify the relative importance of each strand or facet of the standards for testing purposes. This sequential process continues with the development of item specifications, which delineate acceptable item formats, expected cognitive demand levels of items linked to component of the standards, and if items are to be linked directly to objectives within the standards, more general aspects of the standards, or to specific curricular components. Items are developed by following test specifications, and commonly undergo review for clarity, accuracy, potential bias, and alignment with the standards. In many states, items are linked directly to specific performance objectives that comprise the academic standards. Items that pass review are field tested and checked for statistical properties before becoming operational on later test forms.

Test construction activities are vital for test-standard alignment, but are limited in that individuals external to the development process rarely are involved. As is the case in any comprehensive evaluation, it is necessary to obtain feedback from experts outside the system because they can provide test sponsors and developers a much-needed fresh perspective on how tests are working to measure standards. A thorough external review should yield objective summative evaluation information about a test, and substantive formative information about how a test can be improved. Because testing is a continually evolving process (i.e., items are replaced over time and standards are modified on occasion), alignment analysis should not be perceived as a one-time activity, but rather as a critical step in the test evolution loop.

Alignment analysis is not a new process or one germane to standards-based assessment. As long as educators have been linking test items and learning objectives, the need for examining the connection between the two has existed. But with the advent of standards-based reform, a number of comprehensive alignment methods have emerged. The three most commonly employed models include, the Web Alignment Tool (WAT), the Achieve Assessment-to-Standards Model (the Achieve Model) and the Survey of Enacted Curriculum (SEC). These

models build on earlier, basic alignment methods known as matching and rating. Matching involves asking SMEs to choose the objectives from the standards that best fit each test item. SME agreement is indicative of high alignment. An item is considered to be "aligned" with an objective if a large proportion of SMEs match the item to the objective. In rating, SMEs are provided an item and objective connection and asked to judge on a multi-point scale the degree to which the item aligns with the objective. In both matching and rating, SMEs can be asked to gauge alignment based on item and objective content congruence, cognitive demand congruence, or both.

Based on prior research conducted on the 2004 AIMS high school mathematics exam that revealed the advantages of matching over rating, the WAT method was chosen to evaluate the alignment of the 2010 AIMS mathematics exams. The WAT, which primarily is a matching technique, combines both quantitative and qualitative alignment evidence. After SMEs rate the cognitive complexity of both items and objectives, match items to objectives, and record any comments or concerns they have about specific items, their findings are summarized using five criteria: Categorical Congruence, Depth of Knowledge, Range of Knowledge, Balance of Representation and Source of Challenge. Categorical Concurrence refers to the extent to which the standards and an assessment incorporate the same content. Depth of Knowledge indicates whether the assessment requires students to answer items on the test that are at least as challenging as those outlined in the standards. Range of Knowledge is the proportion of performance objectives in the state standards that are measured on the test. Balance of Representation is a measure of item spread across objectives. Finally, Source of Challenge refers to comments reviewers make about items to indicate that they may need revision.

Collecting Alignment Analysis Data

Participants

Twenty-three SMEs reviewed the alignment between AIMS mathematics exams and the Arizona Academic Standard in mathematics over a two-day period (March 4th and 5th, 2010) at Ottawa University in Phoenix, Arizona. Participants were recruited from various regions of the state, including Phoenix, Tucson, Flagstaff, and more rural areas. Table 1 presents the SMEs' background characteristics.

Each SME was asked to serve on one of four committees based on their grade-level experience. A committee consisting of six judges reviewed the grade 3 and 4 exams, while another six individuals evaluated the grade five and six tests. Five SMEs judged the grade seven and eight exams, and six judges reviewed the high school test. As can be seen from Table 1, the SMEs from each committee were very experienced, with all but 5 members having more than ten years teaching experience, and 14 members had more than 20 years of teaching experience. All six of the SMEs on the high school test committee had teacher certifications in mathematics, and 60 percent of the grade 7 and 8 SMEs had mathematics certification. All judges had experience teaching the grade levels for which they judged tests. Most SMEs had advanced degrees, presently were classroom teachers, were women, and were white, though other racial groups were represented among the total group. There was good diversity across SMEs regarding the

location of their schools (urban, suburban, and rural) and the socioeconomic levels of their students' families.

Table 1

Subject Matter Expert (SMEs) Characteristics by Committee

	Committee			
	Grades 3 &4 (*n*=6)	Grades 5 & 6 (*n*=6)	Grades 7-8 (*n*=5)	Grade 10 (*n*=6)
Demographics				
Male		1 (17%)		2 (33%)
African American				1 (17%)
Native American				1 (17%)
White	6 (100%)	6 (100%)	5 (100%)	4 (67%)
Highest Degree				
Bachelor's	2(33%)	2 (33%)	2(40%)	1(17%)
Bachelor's + 72 Credits	1(17%)			
Master's	2(33%)	4 (67%)	3(60%)	4(67%)
Doctorate	1(17%)			1 (17%)
Current Position				
Teacher	4(67%)	6 (100%)	4 (80%)	4 (67%)
Curriculum Specialist	2(33%)			1 (17%)
Math Coach			1(20%)	
School Improvement Director				1(17%)
Match Certified		3 (50%)	3(60%)	6(100%)
Taught Grade of Tests Reviewed	6(100%)	6(100%)	5(100%)	6(100%)
Present School Location				
Urban	2(33%)	1(17%)	3(60%)	1(17%)
Suburban	2(33%)	2(33%)	2(40%)	3(50%)
Rural	2(33%)	3(50%)		2(33%)
Years Teaching Experience				
10 or less years	1(17%)	1(17%)	2 (40%)	1 (17%)
10-20 years	1(17%)	2 (33%)		1 (17%)
Over 20 years	4(67%)	3 (50%)	3 (60%)	4 (67%)

AIMS Mathematics Exams

The AIMS mathematics tests were created to measure the Arizona Academic Standard in mathematics. Committees of Arizona educators developed the multiple-choice test questions to measure specific objectives in the state standard. Separate committees reviewed the items for sufficient match with the objectives and potential bias, and approved items were field tested on prior year exams.

Mathematics items were created to measure objectives in the five strands that comprise the Academic Standard: (1) Number Sense, (2) Data Analysis and Probability, (3) Patterns, Algebra, and Functions, (4) Geometry and Measurement, (5) and Math Structure and Logic. State test specifications stipulated that each test question was to measure one objective from the Academic Standard. Table 5 provides the number of test items per strand according to the state item map or specifications. It can be seen that an ample number of items were included on tests to measure each strand across the grade levels, with Number Sense having the most linked items in the earlier grades and Patterns, Algebra, and Functions, and Geometry and Measurement emphasized more in the upper grades, especially the high school exam.

Procedures

In the morning of the first day, Dr. D'Agostino led a two-hour training workshop in which SMES became acquainted with the WAT computer program, reviewed the Depth of Knowledge (DOK) scoring rubric, and practiced rating the DOK levels of items and objectives. SMEs also practiced matching items to objectives and identifying any Source of Challenge issues. Items and objectives from a sixth grade assessment and standards from another state were used during the training session.

The item and objective DOK, or cognitive demand level, was coded based on Norman Webb's system. SMEs were asked to code each objective and item into one of four levels including: (1) recall; (2) skill/concept; (3) strategic thinking; and (4) extended thinking. After each SME coded the objectives, the committees met separately with an assigned group leader to reach a consensus on the DOK levels of the objectives that comprised their grade level standard. Group leaders orchestrated committee discussion, and then entered the final consensus DOK levels for the objectives into the WAT program. Table 2 presents the number of items and reviewers per test, as well as the interrater reliability coefficients of the SMEs on the objective DOK levels based on their individual ratings. As can be seen, the reliability values are quite strong, with a low of .68 for high school and a high of .89 for grades four and five. In general, the SME agreement on objective DOK classifications was stronger for the younger grade exams.

Following the DOK rating of the objectives, which took about three hours, the SMEs were asked to take the tests, score the DOK level of each item, and match each item to up to three objectives from the standards. If SMEs could not identify a viable match for an item, they were asked to score the item as "uncodable."

Table 2

Test Items, Subject Matter Experts (SMEs), and Reliability by Grade

Grade	Test Items	SMEs	SME Reliability
3	66	6	.83
4	68	6	.89
5	67	6	.89
6	68	6	.83
7	68	5	.70
8	68	5	.70
10	85	6	.68

SMEs also were instructed to record any particular issues they detected with items in the "Source of Challenge" textbox for each item, and to type either a "Yes" or "No" in the same textbox for each item if they felt the item was age appropriate for students or not.

Each SME was asked to work alone and to not consult their group members while coding DOK and matching items to objectives. They entered item ratings and item-objective matches directly into the WAT database software. Once ratings were entered, the software program automatically generated reports on various aspects of content alignment.

Alignment Analysis Results

The WAT program compiles the item and objective DOK scores, the SME matches of items and objectives, and the source of challenge comments to derive alignment values on five dimensions, including:

1. **Categorical Concurrence,** the extent to which the content contained in the standards is assessed. A strand meets this criterion if more than six assessment items target that strand. The value of six items was arbitrarily set by WAT developers as the minimum number of items necessary to compute a reliable strand score.

2. **Depth of Knowledge Consistency**, the degree to which test items require the same complexity of thinking as required by the standards. A strand meets this criterion if more than half of the assessment items are as complex as the objectives they target.
3. **Range of Knowledge Correspondence**, whether the span of knowledge described in a strand corresponds to the span of knowledge required to correctly answer test items. A strand meets this criterion if more than half of the objectives associated with a strand are assessed by at least one item.

4. **Balance of Representation,** the degree to which one objective is given more emphasis on the assessment than another. A strand meets this criterion if, among assessed objectives, similar numbers of items are associated with each objective.

5. **Source of Challenge,** any characteristic of a test item that inhibits its ability to measure the objective of interest. An item is flagged as having a source of challenge issue if reviewers thought that the item was unclear, confusing, or had some other issue that prevented it from measuring a performance objective well.

Arguably, Categorical Concurrence is the primary dimension of alignment because it addresses the degree to which test items align with the objectives in the state standards. For this reason, it is important to understand the computational procedure utilized by the WAT program to compute Categorical Concurrence. Table 3 provides the Number Sense Concurrence data for a sample test containing 22 items.

Table 3

Sample Categorical Concurrence data

Item #	Item Map Strand	Objective Strand Number of SME Matches														*Strand 1* WAT Method	Hits	Hits WAT Method
1	1	1	1	1	1	1	1	1	1	1	1	1	4			.92	1	.92
2	1	1	1	1	1	1	1	1	1	1	3					.75	1	.75
3	3	1	3	3	3	3	3	3	3	3	3	3	3	3		.08		
4	3	1	1	3	3	3	3	3	3							.17		
5	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	.08		
6	4	2	4	4	4	4	4	4	4	4	4	4				0.0		
7	3	1	1	1	1	1	1	1	1	1	1	3	3	3	3	.83		
8	4	4	4	4	4	4	4	4	4	4	4					0.0		
9	4	4	4	4	4	4	4	4	4	4	4	4				0.0		
10	1	1	1	1	1	1	1	1	1	1	1	1	2	3	3	.92	1	.92
11	4	4	4	4	4	4	4	4	4	4	4	4				0.0		
12	5	5	5	5	5	5	5	5	5	5						0.0		
13	2	1	1	1	1	1	1	1	1	1	1	1	1	3		1.0		
14	2	1	2	2												.08		
15	2	2	3	3	3	3	3	3	3	3	3	3	3	3		0.0		
16	2	2	2	2	2	2	2	2	2	2	3					0.0		
17	2	1	1	1	1	1	1	1	1	1	2					.83		
18	1	1	1	2	2	2	2	2	2	2	2	2	2	2		.17		.17
19	2	2	2	2	2	2	2	2	2	2	2	2	2	2		0.0		
20	3	2	2	2	2	2	2	2	2	2	2	2	2	2		0.0		
21	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0.0		
22	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0.0		0.0
															Totals	5.83	3	2.76

The strand from the state mathematics standard that each item was linked to according to the state item maps is presented in the column to the right of the item number column. As can be seen, for example, there were five items linked to Strand 1 (Number Sense) on the exam. The fourteen columns to the right of the strand number column present the strand number for the objectives that SMEs matched to each item. Each column represents a unique SME match. For instance, SMEs made eleven matches between an objective in Strand 1 and Item 1, and one SME matched an objective from Strand 4 to that item. In some cases, such as for Item 5, there were

more objectives matched to an item than there were SMEs (12) who reviewed the test, because SMEs could match up to three objectives to an item. In other cases, such as Item 14, fewer than 12 matches were produced, revealing that most SMEs considered the item "uncodable" or not matched to any objective.

The WAT program computes Concurrence for a strand by first tallying the number of matched objectives from the strand for each item, and then dividing that total by the number of SMEs. For example, eleven Strand 1 matches were made for Item 11. Dividing that total by 12 produces a value of 0.92. On Item 3, only one match to an objective in Strand 1 was rendered by an SME, so the Strand 1 Concurrence for that item was 0.08 (1/12). The logic of this approach is that Strand 1 is given 0.92 points for Item 1 because 11 of 12 judges matched the item to an objective from the strand. The Concurrence values for each item on a strand are totaled to produce the final Concurrence value for the strand. If the final summed value for a strand is less than 6, Concurrence is not met, because 6 items are deemed necessary to produce a reliable strand score.

Notice, however, that Strand 1 received 0.83 points for Item 7 and 1.0 full point for Item 13 even though both items were mapped to other strands. Thus, the WAT Concurrence values reflect the alignment of items to strands without any consideration for the item maps. The method reveals how items can be linked to strands according to SMEs, but it does not provide a verification check on the state item maps. Furthermore, notice that more than one strand often receives multiple Concurrence points for each item. For example, SMEs matched Item 11 only to objectives in Strand 4, so only that strand received alignment points for the item, but Strands 1 and 3 both received Concurrence points for Item 7. According to state test specifications, however, each item is to be linked to one objective, so it is not possible for an item to be used to compute multiple strand scores.

Another concern with relying on the WAT method as the sole indicator relates to the criterion rule of at least six items required for strand Concurrence. Some AIMS tests were not designed to measure certain strands with more than ten items, which would greatly reduce the likelihood that Concurrence would be met in those strands. Arizona does not rely on the strand scores for high-stakes purposes, and does not attempt to set cut-scores by strand level, so the WAT criterion of at least six matched items does not apply in these cases. Also, because the WAT method does not evaluate the state item maps, a strand can be deemed aligned because six or more matches were made by SMEs, even if SMEs did not match the same items used by the state to compute the strand scores to that strand. Thus, the WAT can produce misinformation about the validity of the strand scores.

To address these limitations with the WAT approach, and to focus on evaluating the state item maps, alternative Concurrence indices were computed to examine the alignment of the AIMS tests. One simple alternative, which will be called the Hits WAT method, involves following the same computational procedures as the WAT, but counting strand points only for items linked to a given strand as specified by the item maps. These values are represented in the far right column of Table 3 for Strand 1 on the sixth-grade mathematics exam. As can be seen in the table, Strand 1 received the WAT method Concurrence points for items 1, 2, 10, 18 and 22 on the sample test, but not for the items that were not linked to the Strand per the item map.

Another approach, termed Hits, was employed that involved tallying the number of items linked to each strand according to the item maps that received at least half the number of matches as there were SMEs that judged the exam. The table reveals that Items 1, 2, and 10 from the sample test were deemed Hits because there were at least six matches for each item to Strand 1 objectives by the 12 SMEs. Items 18 and 22 were not deemed Hits because each item received fewer than six Strand 1 matches. Instead of expecting each strand to contain at least six item hits, a strand was considered aligned in terms of Concurrence if the majority of items linked to the strand according to the item map were hits. For example, the sample test met Concurrence for Strand 1 because three of the five linked items were Hits.

The WAT dimensions and the alternative methods to evaluate Categorical Concurrence address different alignment facets, and taken together, provide comprehensive feedback about the congruence between test items and state standards.

Alignment Results for 2010 AIMS Mathematics

Table 4 below presents the summary WAT results for each 2010 AIMS mathematics test by strand. Overall, the alignment results reveal that AIMS is very well aligned with the state academic mathematics standards, with strengths in Categorical Concurrence, Range of Knowledge, and Balance of Representation. As can be seen from the table, there was but one strand on one grade-level exam in which SMEs did not match at least six items, which was Number Sense on the high school test. Note, however, that by design, only five items were included on the high school test to measure that strand, so the "No" value likely resulted from test design rather than from lack of alignment.

To meet the Range of Knowledge criterion, more than half of the objectives associated with a strand must be assessed with at least one item. Alignment in this dimension is considered "Weak" if slightly less than half of the objectives from a strand were matched with at least one item. As can be seen from the table, in only two cases, high school Number Sense and grade 8 Math Structure and Logic, did the AIMS tests not meet the criterion. There were two tests (grades five and seven) judged to be "Weak" for Math Structure and Logic. Grade four Data Analysis and Probability also had a "Weak" rating. Range of Knowledge was met for all other strands and tests.

In Balance of Representation, which is met if, among assessed objectives, similar numbers of items are associated with each objective, grade five Math Structure and Logic was deemed to be "Weak," but in all other cases, the AIMS tests met this criterion. One area that apparently can be improved is the Range and Balance of items across objectives for Math Structure and Logic, particularly in grade. All other tests sufficiently measured a breadth of objectives within and across strands.

The findings for Depth of Knowledge were not as strong, but overall were rather positive. The primary lack of alignment in the DOK ratings for objectives and matched items was for Data Analysis and Probability (one of seven "Yes" ratings) and for Math Structure and Logic (two of seven "Yes" scores). A more detailed analysis of the data showed that SMEs tended to rate objectives as having higher DOK levels than the matched items. More objectives in those strands

Table 4

WAT Ratings by Strand and Grade Level

	Categorical Concurrence				
	Number Sense	Data Analysis, Probability	Patterns, Algebra, Functions	Geometry, Measurement	Math Structure, Logic
Grade 3	Yes	Yes	Yes	Yes	Yes
Grade 4	Yes	Yes	Yes	Yes	Yes
Grade 5	Yes	Yes	Yes	Yes	Yes
Grade 6	Yes	Yes	Yes	Yes	Yes
Grade 7	Yes	Yes	Yes	Yes	Yes
Grade 8	Yes	Yes	Yes	Yes	Yes
Grade 10	No	Yes	Yes	Yes	Yes

	Depth of Knowledge Consistency				
	Number Sense	Data Analysis, Probability	Patterns, Algebra, Functions	Geometry, Measurement	Math Structure, Logic
Grade 3	Yes	No	Yes	Yes	Yes
Grade 4	Yes	Weak	Yes	Yes	No
Grade 5	Yes	No	No	Yes	Weak
Grade 6	Yes	Weak	Weak	Yes	Yes
Grade 7	Yes	Weak	Yes	Yes	No
Grade 8	Yes	No	Yes	Yes	Weak
Grade 10	Yes	Yes	Yes	Yes	No

	Range of Knowledge Correspondence				
	Number Sense	Data Analysis, Probability	Patterns, Algebra, Functions	Geometry, Measurement	Math Structure, Logic
Grade 3	Yes	Yes	Yes	Yes	Yes
Grade 4	Yes	Weak	Yes	Yes	Yes
Grade 5	Yes	Yes	Yes	Yes	Weak
Grade 6	Yes	Yes	Yes	Yes	Yes
Grade 7	Yes	Yes	Yes	Yes	Weak
Grade 8	Yes	Yes	Yes	Yes	No
Grade 10	No	Yes	Yes	Yes	Yes

	Balance of Representation				
	Number Sense	Data Analysis, Probability	Patterns, Algebra, Functions	Geometry, Measurement	Math Structure, Logic
Grade 3	Yes	Yes	Yes	Yes	Yes
Grade 4	Yes	Yes	Yes	Yes	Yes
Grade 5	Yes	Yes	Yes	Yes	Weak
Grade 6	Yes	Yes	Yes	Yes	Yes
Grade 7	Yes	Yes	Yes	Yes	Yes
Grade 8	Yes	Yes	Yes	Yes	Yes
Grade 10	Yes	Yes	Yes	Yes	Yes

Table 5

WAT and Alternative Categorical Concurrence by Strand and Grade Level

	Categorical Concurrence				
	Number Sense	Data Analysis, Probability	Patterns, Algebra, Functions	Geometry, Measurement	Math Structure, Logic
Grade 3					
Item Map	28	8	11	12	7
Hits	26 (Yes)	8 (Yes)	10 (Yes)	11 (Yes)	6(Yes)
WAT CC	35.67 (Yes)	10.83 (Yes)	12.17 (Yes)	11.00 (Yes)	12.33 (Yes)
Hits WAT CC	26.65	9.17	8.67	10.83	6.16
Grade 4					
Item Map	27	8	12	12	9
Hits	26 (Yes)	7 (Yes)	11 (Yes)	11 (Yes)	7 (Yes)
WAT CC	29.83(Yes)	9.67 (Yes)	13.33(Yes)	10.50(Yes	12.50(Yes)
Hits WAT CC	25.34	7.17	11.50	9.17	7.17
Grade 5					
Item Map	25	12	11	10	9
Hits	24 (Yes)	12 (Yes)	8 (Yes)	10 (Yes)	7 (Yes)
WAT CC	26.33(Yes)	14.17(Yes)	9.33(Yes)	11.33(Yes)	12.00(Yes)
Hits WAT CC	24.83	12.17	7.67	10.17	6.67
Grade 6					
Item Map	23	12	11	13	9
Hits	21 (Yes)	12 (Yes)	11 (Yes)	13 (Yes)	7 (Yes)
WAT CC	25.17(Yes)	14.67 (Yes)	14.00 (Yes)	13.50 (Yes)	12.17 (Yes)
Hits WAT CC	22.82	12.83	11.67	12.50	6.83
Grade 7					
Item Map	17	13	13	15	10
Hits	15 (Yes)	11 (Yes)	12 (Yes)	15 (Yes)	4 (No)
WAT CC	19.00(Yes)	13.40(Yes)	13.20(Yes)	16.20(Yes)	8.00(Yes)
Hits WAT CC	15.40	10.60	13.00	13.40	5.40
Grade 8					
Item Map	12	12	18	16	10
Hits	9 (Yes)	11 (Yes)	16 (Yes)	15 (Yes)	5 (Yes)
WAT CC	10.40(Yes)	12.80 (Yes)	21.00 (Yes)	17.20(Yes)	9.00(Yes)
Hits WAT CC	8.80	11.00	16.80	15.00	6.40
Grade 10					
Item Map	5	12	28	28	12
Hits	3 (Yes)	12 (Yes)	28 (Yes)	27(Yes)	11(Yes)
WAT CC	5.00 (No)	15.67(Yes)	37.83(Yes)	29.33(Yes)	18.17(Yes)
Hits WAT CC	3.67	11.50	30.68	26.99	11.17

Note. For each grade, Item Map indicates the number of items assigned to each strand of the standard, Hits indicates the number of items for which at least 50% of SMEs matched the items to an objective from the same strand as stipulated by the item maps, WAT CC is Categorical Concurrence as calculated by the WAT program, and Hits WAT CC is the number of items SMEs matched to the strand as stipulated by the item maps using the WAT calculation method. A (Yes) next to Hits indicates that at least half of the items linked to the strand as stipulated by the item maps were matched to the strand by at least 50% of the SMEs. A (Yes) next to the WAT CC values indicates that at least 6 items were matched to the strand by SMEs.

require students to perform high-level mental skills such as "creating" or "synthesizing," which likely could be more suitably measured with constructed response items. It seems that matched items required students to perform mental activities that would be foundational for more advanced skills. Depth of Knowledge was strong for the other three strands—Number Sense, Patterns, Algebra, and Functions, and Geometry and Measurement.

The results from the alternative measures of Concurrence, WAT Categorical Concurrence scores, and items per strand according to item maps are presented in Table 5. As can be seen from the table, the alternative measures mostly reflected the initial Categorical Concurrence results as computed by the WAT method. Thus, SMEs not only tended to match at least six items to strands, they also tended to match the items to strands as specified by the test item maps. In only two cases were the results of the two methods discordant. In grade seven Math Structure and Logic, SMEs matched eight items to the strand, but only four (Hits method) or 5.40 (Hits WAT method) of those items were designed to measure the strand. Categorical Concurrence based on the WAT method was not met for high school Number Sense, yet three of the five items designed to measure the strand were matched to it by SMEs. Clearly, SMEs matched verified the validity of test item maps by matching items to objectives according to test specifications.

Table 6 provides a list of items flagged by at least two reviewers that had Source of Challenge issues. SMEs were instructed that providing comments in this area was optional. Note that the committee that reviewed the grades three and four exams, and the committee that reviewed the grade seven and eight tests, tended to utilize this option to a greater extent. Table 7 includes all unedited Source of Challenge comments per test and item. Each row in the tables represents one SME's comment. If an item number is missing from the table, no SME commented about the item. Most Source of Challenge comments focused on SMEs failing to identify a good matching objective, incorrect item answers or distracters, confusing prompts, or issues with the visual display of items.

Table 6

Grade	Items
3	5, 16, 21, 37, 38, 79
4	5, 7, 9, 19, 23, 31, 37, 39, 77
5	73
6	83
7	1, 2, 8, 25, 26, 28, 36, 51, 57, 62, 67
8	15, 70
10	None

Table 7

Source of Challenge by Item, Grade 3

Item	*Source of Challenge*
5	Double digit multiplication??
5	This problem requires students to multiply 20x3, which is not a 3rd grade standard. 3rd graders only have to multiply facts through 10's according to our standard. This problem is not grade level appropriate.
6	if this is not meant for po 1.2.7 (commutative property) the formula that matched the problem should be given
12	This problem must not be coded under 3.1.2 because it is not a numerical pattern. It is a geometrical pattern, so it must be coded as a 4.1.1
15	3/8 is not a benchmark fraction
16	If this item is not meant for the above PO it is uncodable. If meant for 4.4.1 it is not across months as specified in the PO
16	This does not follow the specifications within the PO. It should be elapsed time across months.
21	'nearer'generally is not the word used in this situation - 'nearest' is what 3rd graders normally would use. Additionally, the idea of rounding is not a 3rd grade skill.
21	Item should say "Stick on digits" instead of stick on numbers. Can be confusing and is not correct term.
24	Is there great enough difference between the objects in choice A (block and shoe box) to discount similarity at the 3rd grade level?
37	You don't start measuring in the middle of a ruler.
37	Measuring in the middle of a ruler makes no sense.
38	Solving problems with elapsed time is a 6th grade objective. No Clock is used per 3rd grade standard.
38	Does NOT align! PO states must use clock for elapsed time. Also PO is not problem solving using elapsed time - it is only determining elapsed time.
47	Use the term "greater" for numbers, not "higher".
49	PO states benchmark fractions, not necessarily fractions with like denominators.
52	"Nuts as a favorite choice for boys? Seriously? Are they to imply that boys and girls are "students"? Same verbiage should be in question and stem.
55	Should not multiply by a factor greater than 10. Also 3rd graders do not add money.

59	counting frame representation not a universal model/representation
70	You can't make that exact figure. It dosen't say similar.
79	name in data for problem section (Jenny)does not match name in problem (Jenna)
79	Stem says Jenny, question says Jenna.
79	Name in graphic doesn't match name in stem
80	name in data for problem section (Jenny)does not match name in problem (Jenna)

Source of Challenge by Item, Grade 4

Item	*Source of Challenge*
5	If not 1.1.2 this should be uncodable as it does not correspond to any other PO.
5	Doesn't meet PO. Just a portion of it.
5	doesn't match any P.O. at this grade
5	This item is difficult to align to a PO. The item really has students convert money, which is not specifically in any particular PO. The measurement PO about conversions or an estimation PO seemed to be the closest fit.
7	Venn diagrams are not included in gr. 4 standard.
7	Venn not mentioned in 4th grade
7	Not a PO
7	Venn Diagrams used?
7	No Venn diagrams at 4th grade objectives
7	Venn diagrams are not in the 4th grade standard!
9	You wouldn't teach it like this.
9	Item really involves students dividing using 3 digit divisors which is beyond 4th grade
9	If students solve this equation using inverse operations, then they are required to divide with a three digit divisor....not grade level appropriate.
18	Question in problem relates to scores of both girls yet 2 of 4 responses (including the correct response) only relate to scores of 1.
19	equivalent fractions not part of 4th gr. standard
19	Equivalency of fractions not mentioned in standard
19	Doesn't match PO
19	Not found in standards
19	Equivalent fractions not at this grade level
19	There are no equivalent fractions in the 4th grade standard.
23	Which part is painted????
23	Could be either of two possible answers because the graphic doesn't show which part of the fence is painted

23	This item needs to specify what color is being painted...the white or the gray? Depending on what color the students interpret as the paint, there could possibly be 2 correct answers here.
31	Confusing graphic.
31	This item has a very complex context. This context is not grade level appropriate.
37	loosely connected to the PO
37	Fact families are specifically placed in the 3rd grade standard. This item can be tied to the PO in 4th grade about fluency of multiplication facts, but this is truly a 3rd grade item.
39	Comparing fractions with unlike denominators is not a PO.
39	comparing fractions with unlike denominators not at this grade level
43	If students solve this item by computing, then this item is outside of the 4th grade standard as it requires multi-digit by two-digit multiplication. It should be multi-digit by one-digit or two-digit by two-digit as specified in the 4th grade standard.
45	Second standard should be 45.2.3. Error in computer.
53	At this grade level, students should not be expected to know that T = Total. Item should include information that states T = Total.
54	This conversion is not 4th grade appropriate...much too difficult.
77	Too difficult combining 2 PO's for fourth grade.
77	combining estimation with probability is above this grade level

Source of Challenge by Item, Grade 5

Item	*Source of Challenge*
1	Metric/US standard combined
9	The number set and operations seem low for the 5th grade objectives
23	Trapezoid - compound should state figure, not trapezoid
27	Formal algebraic notation not in the 5th grade standards. This is not in the standards until grade 6
28	Formal algebraic notation not in the 5th grade standards. This is not in the standards until grade 6
35	Possibly below grade level.
43	This is multiplication.
69	Too low for grade level
73	Is formal algebraic notation appropriate at this level?
73	In 5th grade?

Source of Challenge by Item, Grade 6

Item	*Source of Challenge*
1	A little easy for this PO
35	Not really something a 6th grader would do.
38	Too low for 6th grade.
56	Most obvious error not one of the answer choices
77	Doesn't really fit our standards.
83	Doesn't really fit the POs
83	Uses an equation rather than an expression. Testes the commutative property, but with decimals rather than whole numbers.
83	Identifying communitative property; not whole numbers

Source of Challenge by Item, Grade 7

Item	*Source of Challenge*
1	no justification - only identification
1	not a complete match
2	Couldn't identify a PO
2	No labels shown, does not fit standard
5	shortest path is not a real e or h
8	is this a two -step?
8	Wording of stem is hard to understand. Take out 2 fish, draw next fish what is chance of third fish...do you draw two more? I was confused.
10	2nd sentence is extraneous
12	1st sentence is extraneous
14	3 of the 4 answers have the word 'number' student may elimate those 3 based on that
15	formula sheet will confuse students possibly; this question along with two others will give answers away
16	should not be next to previous questions as table gives the answer away
18	not grade appropriate; numbers are too easy since integers and fractions should be a focus
20	no comparison of probability
23	not contextual
24	no create
25	The word 'steady' is in the correct answer but the decline is not steady (which I interpret as being the same amount) from month to month
25	no table or data set; what is steadily?
26	PO says perimeter and area but not volume
26	This is just a formula problem comparing the two results. I couldn't link to any PO
26	no calculation of volume in po
27	too easy for grade level
28	po does not ask to compute

28	The answer to this question can be found by going to question 15 in same section. Maybe move one of the problems to a different section
32	no tables shown
36	Confusing problem - maybe the wording??
36	No probability. there could be than 1 answer
41	not a good match
45	weak fit;
46	no create
50	not a good fit as this question basically is a compute slope
51	The word 'closest' indicates an approximation by the answer is right on.
51	no area
54	not grade appropriate
56	Same problem as question 39, first answer is correct on both problems. Perhaps change order of choices so students have to check more than one answer.
57	Precursor to standard??
57	not grade appropriate; multiple answers; does not fit the po. a really really bad question !!
62	2 variables
62	Precursor to being able to solve?
63	shortest path; not a good fit
64	no table
67	Precursor? Doesn't test the PO
67	not grade appropriate
81	not grade appropriate
83	The total of three of the choices equals the same - should all 4 equal the same? I think students may choose that answer just because the total is different.
84	combination/permutation problem; not a po
85	not grade appropriate

Source of Challenge by Item, Grade 8

Item	*Source of Challenge*
14	Is this grade appropriate?
15	Is this problem a precursor to the objective?
15	not grade appropriate
18	Not compound
23	not grade appropriate; a gotcha question. 4 a a choice does not mean students understand
26	no model
30	not grade appropriate
31	7th grade standard
44	wow!
46	poor distractors
48	Negative exponents not 8th grade standard
51	I think the correct answer should say "people who own neither a cat or dog"
56	poor fit
57	too easy for 8th grade
59	It was just picking an algorithm to use - could it be a precursor to this objective?
60	Just have to pick out an algorithm t use...is this a precursor problem?
61	7th grade
68	no proportional reasoning
70	This problem was confusing - change in distance from home??
70	not slope
70	bias? Will "running" errands imply "running" for some students? Then the graph is confusing.
79	7th grade standard
81	not grade level
84	Problem extended after identifying.
87	not sure about the answer

Source of Challenge by Item, Grade 10

Item	*Source of Challenge*
3	Not really finding the midpoint
5	Choice B and D could be read as equivalent answers with slightly different wording.
16	This should be written in function notation.
46	The distractors on this question significantly lower the DOK level. Only one is even viable - choice A. Another distractor could include 4 first followed by it's logical conclusion and then make a mistake in the order. Another choice that begins with 1, 4 and then makes a mistake in order, etc.
48	This item tests an objective that is in the 4th grade and continuing through 6th grade standard. There does not exist an objective at the HS level that can be connected to this item. It is a more appropriate item for 4th, 5th or 6th grade.
53	Isn't this a 7th or 8th grade PO?

Conclusion and Recommendations

Over the two-day alignment period, SMEs provided valuable information regarding the alignment of the 2010 AIMS mathematics exams. They worked diligently to render accurate and detailed information on the tests and standards. Not only can their feedback be used to judge the tests, but it can be used to improve future, and possibly academic standards. It is highly recommended that ADE staff review the Source of Challenge comments in Table 7 to identify any items for potential modification, particularly items flagged by multiple SMEs. In some cases, Source of Challenge problems might reveal that objectives need revision.

Though the WAT data is very rich and results vary considerably across grade levels, there are some general trends apparent in the tables presented in this report. First, the weakest dimension of WAT alignment was Depth of Knowledge, specifically in Data Analysis and Probability and Math Structure and Logic. In most cases, SMEs rated the objectives from those strands as requiring higher DOK levels, but judged the items they matched to those objectives to be lower DOK levels. This finding likely is due to the nature of the standard—many of the objectives from those strands require students to construct or create a response which is simply not possible with multiple-choice items. The matched items do, however, require the skills necessary to achieve the higher levels of cognitive demand, so in a sense, are partially representative of those objectives. The Depth of Knowledge results for the other three strands of the standard were positive overall.

The strongest WAT dimension of alignment was Categorical Concurrence. In only once case, high school number sense, was Concurrence deemed to not be sufficient, and in the one "No" case, there was a small number of items included on the test to measure the dimension. The alternative methods of Categorical Concurrence revealed that SMEs tended to match items with objectives that coincided with the test specification, which provided validity evidence for the item maps and strand scores.

The evidence for Range of Knowledge and Balance of Representation was mostly favorable, with some minor weaknesses for Range of Knowledge in Math Structure and Logic. Indeed, Math Structure and Logic was the strand with the most issues. Often items created to measure logic and structure are contextualized within an algebraic or geometric problem, which creates ambiguity for SMEs (hence, they tend to match those items to objectives from algebra or geometry strands).

The alignment between the AIMS tests and Arizona academic mathematics standard was judged to be strong by SMEs who participated in this study. Overall, the WAT scores provided solid content validation evidence that supports the notion that AIMS mathematics test questions measure the academic standard. All of the grade-level tests were judged to be aligned with the grade specific objectives, indicated by mostly "Yes" values across the strands and WAT dimensions for each exam. All mathematics tests seem to be working effectively to measure students' attainments of the standard. Feedback from outside experts is necessary to make AIMS the best testing programs possible for Arizona schools.