Alignment Analysis of Arizona Academic Standards and Assessments

Jerome V. D'Agostino, Megan E. Welsh, & Adriana D. Cimetta

University of Arizona

September 2005

Alignment Analysis of Arizona Academic Standards and Assessments

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. From July 25th to July 27th, 2005, subject matter experts (SMEs) reviewed the congruence between the Arizona Academic Standards and the 2005 Arizona Instrument to Measure Standards (AIMS) mathematics and reading assessments for grades three through eight and high school, and Arizona Instrument to Measure Standards—Alternative (AIMS—A) Levels I and II in mathematics and reading. The alignment study was conducted by researchers from the University of Arizona, Department of Educational Psychology. 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 2005 AIMS and AIMS-A 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.

The next section provides (1) information on how SMEs were recruited and their characteristics, (2) information about the tests that were reviewed and the WAT alignment tool, and (3) details on the procedures followed to collect alignment evidence.

Collecting Alignment Analysis Data

Participants

Fifty-seven SMEs reviewed the alignment between AIMS items and the Arizona standards over a three-day period (July 25th to July 27th, 2005). The first day was spent reviewing mathematics items (with 24 SMEs participating), the second day was devoted to reading (with 22 SMEs participating), and special education items (in mathematics and reading) were addressed on the final day (with 29 SMEs participating).[1] Participants were recruited from throughout the state.

Table 1 presents reviewers' background characteristics. All SMEs worked in Arizona school districts, either as classroom teachers (between 32 and 71 percent of SMEs), special education teachers (between 10 and 39 percent), school district administrators (between 10 and 14 percent), or in some other capacity, such as a school psychologist, a school counselor, or a Title I teacher (between 10 and 19 percent). Although several SMEs no longer worked in the classroom, all had

[1] Some SMEs participated on more than one day or reviewed documents at multiple grade levels. The numbers used in the remainder of this section represent duplicate counts (e.g., SMEs who rated both a reading and a math test are counted twice, as are SMEs who rated math tests at more than one grade level.)

extensive teaching experience. Nearly half had at least 20 years of teaching experience and between 18 and 28 percent (depending on the subject area) taught in Arizona for 20 years or more.

Table 1

Subject Matter Expert (SMEs) Characteristics by Subject Area

	Reading (n=22)	Math (n=24)	Special Education (n=29)
Demographics			
Male	19.0%	19.0%	10.7%
White	85.7%	70.0%	77.8%
Hispanic	0.0%	5.0%	0.0%
African American	4.8%	10.0%	14.8%
Native American	0.0%	10.0%	3.7%
Asian	0.0%	0.0%	0.0%
Mixed Heritage	9.5%	5.0%	3.7%
Highest Degree			
Bachelor's	36.4%	25.0%	32.1%
Master's	54.5%	70.0%	64.3%
Doctorate	9.1%	5.0%	3.6%
Current Position			
Classroom	52.4%	71.4%	32.1%
Special Education	14.3%	9.5%	39.3%
Administrator	14.3%	9.5%	14.3%
Other	19.0%	9.5%	14.3%
Arizona teaching experience			
0 years	0.0%	4.8%	3.8%
1-4 years	18.2%	23.8%	15.4%
5-9 years	36.4%	19.0%	15.4%
10-14 years	9.1%	9.5%	15.4%
15-19 years	18.2%	14.3%	26.9%
>=20 years	18.2%	28.6%	23.1%
Total teaching experience			
0 years	0.0%	0.0%	0.0%
1-4 years	9.1%	4.8%	6.9%
5-9 years	13.6%	14.3%	13.8%
10-14 years	13.6%	9.5%	10.3%
15-19 years	18.2%	28.6%	20.7%
>=20 years	45.5%	42.9%	48.3%

SMEs were also highly educated. More than half had a master's degree or a doctorate in education (one had an MFA). Similar to the gender and ethnicity of Arizona educators, most reviewers were female and Caucasian.

Instruments

Eighteen tests were reviewed for this study, including 14 AIMS tests (in reading and mathematics, administered to grades 3-8 and high school) and four special education tests (AIMS-A reading Levels I and II, and AIMS-A math Levels I and II). According to the *Alternate Assessment Manual for the Arizona Student Achievement Program*, the Level I tests are designed for students with significant cognitive disabilities who are working towards proficiency on the state's functional and kindergarten level standards. Level II tests are administered to students who have met or exceeded proficiency on the functional and kindergarten standards and who are working toward proficiency on the articulated standards at 1st through 3rd grade.

State standards were also reviewed by grade level and subject area, totaling 28 separate standards documents. The Arizona state standards contain three levels of detail. The strands comprise the most global level and refer to broad skills, such as number sense in mathematics and comprehending literary text in reading. Within each strand, concepts break strands into finer levels of detail, and performance objectives are organized within concepts providing the most in depth descriptions of student expectations. For the purposes of this study, reviewers aligned test items to performance objectives, which are included in the binders appended to this report.

Procedures

SMEs were assigned to rate a particular grade level and subject area in accordance with their area of expertise. In some cases, SMEs were asked to rate an adjacent grade level if they had experience teaching that grade (e.g., an SME who taught 4^{th} grade might have reviewed the 5^{th} test in a given subject) when the number of SMEs varied across grade levels. Reviewers who finished early were also asked to rate adjacent grade levels. Table 2 presents the number of SMEs that reviewed standards and assessments at each grade level/content area.

Data collection followed a similar format on each of the three days. The day started with a one and a half hour training session during which SMEs learned to rate performance objectives and test items according to the cognitive complexity required. Cognitive complexity was coded according to Webb's (1999) Depth of Knowledge levels (DOK). These include four categories: recall, application of a skill or concept, strategic thinking (requiring reasoning, multiple steps, or more than one possible answer), and extended thinking (requiring an investigation, time to plan and carry out a complex task).

DOK training consisted of independently reading through detailed definitions of each DOK level and coding practice objectives and test items by DOK level, followed by a debriefing session in which SMEs discussed how they arrived at their specified DOK level. The training utilized the WAT training materials, including DOK definitions, practice items and practice objectives.

Upon completing DOK training, SMEs rated each performance objective included in the Arizona standards independently. They entered their ratings into the WAT database software using notebooks computers.

Table 2

Number of Subject Matter Experts (SMEs) by Grade Level and Subject Area

	Reading	Math	Special Education Reading	Special Education Math
Grade 3	5	4	4	4
Grade 4	6	4	4	4
Grade 5	4	4	4	5
Grade 6	5	5	4	4
Grade 7	6	4	4	4
Grade 8	6	4	7	4
Grade 10	7	7	7	4

After all SMEs finished rating a grade level, their ratings were reviewed by the research team for discrepant DOK ratings. For objectives with discrepant DOK ratings, reviewers were asked to discuss the objectives and come to consensus on the DOK level of those objectives. Two events flagged such discussions: (1) when the DOK ratings were evenly split among reviewers (i.e., two reviewers coded an objective as recall, and two reviews coded an objective as application of skill/concept, or (2) when DOK ratings were more than one category apart (e.g., one reviewer coded an objective as recall, and another coded it as strategic thinking). Once SMEs agreed on the ratings for discrepant objectives, the consensus DOK level ratings were entered into the WAT database. For objectives that did not require consensus discussions, the majority DOK level ratings were entered.

When the DOK level coding for objectives was completed, SMEs participated in a second training designed to help them think through how to match items to objectives. During this half-hour training, SMEs were asked to identify the content and intellectual skill required by ten performance objectives culled from the Arizona standards and five items from expired AIMS tests. After identifying both the content and intellectual skill required by the items and objectives, they independently practiced matching items to objectives based on both characteristics. SMEs were allowed to match items to as many as three objectives, and were instructed to identify the best matching objective as the primary objective and (when appropriate) other objectives as secondary and/or tertiary matches. Finally, SMEs were instructed to consider the entire item (stem, response options and any supporting material) and both the performance objective and it's overarching concept into account when matching. SMEs discussed their rationale for matching items to objectives after they completed the training exercise independently.

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 are entered, the software program automatically generated reports on various aspects of content alignment. These reports are presented in the next section and in the appended binders.

Alignment Analysis Results

Alignment results were tabulated using two methods. First, SMEs were asked to share their overall impressions of the alignment between assessments and standards on a feedback form distributed at the end of each review. Second, several alignment measures were created from SME reviews by the WAT system. Feedback form results will be discussed first followed by the alignment measures.

Feedback form information is presented in Table 3. Reviewers agreed that test items addressed the most important content for a given grade level and subject area between 41 and 86 percent of the time, with much higher agreement rates for the general AIMS tests than for the special education assessments (AIMS-A). Reviewers were more consistent in their assertion that the assessments addressed the most important intellectual skills (or performance levels), with between 60 and 75 percent of reviewers agreeing. In addition, reviewers tended to report that the content assessed was in line with what they expected, with only 6 to 25 percent of SMEs reporting content missing from the test. Finally, most reviewers reported that the alignment between AIMS and the Arizona standards was either "acceptable" or "needs slight improvement" (82 percent of SMEs in reading and 100 percent of SMEs in math) while nearly half of AIMS-A (special education test) reviewers reported that the assessment "needs major improvement" or is "not aligned in any way."

We used the data collected by the WAT system to generate the following five content alignment measures:

1. **Categorical Concurrence,** the extent to which the content contained in the standards is assessed. A strand meets this criterion if more than five assessment items target that strand.

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.

These measures address different alignment facets and, taken together, provide comprehensive feedback about the congruence between test items and state standards. Tables 4 through 11 present the results on each dimension by test.

Table 3

SME's Overall Impressions of Alignment

Question	Reading (n=22)	Math (n=24)	Special Education (n=28)
Did the items cover the most important topics?			
No	28.6%	14.3%	59.3%
Yes	71.4%	85.7%	40.7%
Did the items cover the most important performance levels?			
No	25.0%	40.0%	37.0%
Yes	75.0%	60.0%	63.0%
Was there content you expected to be assessed that was not included in the test?			
No	75.0%	93.3%	71.4%
Yes	25.0%	6.7%	21.4%
N/A	0.0%	0.0%	7.1%
What was your general opinion of the alignment between the standards and assessment?			
Perfect alignment	0.0%	0.0%	0.0%
Acceptable alignment	50.0%	70.8%	10.7%
Needs slight improvement	31.8%	29.2%	32.1%
Needs major improvement	18.2%	0.0%	46.4%
Not aligned in any way	0.0%	0.0%	10.7%

Reading

At the most basic level, reading items appear to be aligned with state standards. They address the same content outlined by the standards, as indicated by "yes" ratings on all categorical concurrence measures. In general, the intellectual skills required by test items are also similar to those required by the state standards, as indicated by 17 "yes" ratings out of a possible 21 ratings for depth of knowledge consistency (Table 4). However, depth of knowledge consistency was judged insufficient at two grade levels (grades 3 and 8) for the Comprehending Literary Text strand and was rated "weak" at grade 6, both for Comprehending Literary Text and for Comprehending Informational Text. One example of a weak depth of knowledge consistency rating is associated with the third grade objective, "distinguish between/among fiction, nonfiction, poetry, plays and narratives, using knowledge and structural elements." This objective was rated as a level 2 DOK, application of a skill or concept, by SMEs while the item that measured this skill was rated as only requiring recall, DOK level 1. While the item in question did require students to indicate whether a reading selection was a play, poetry, fiction or nonfiction, the selection was easily identifiable. It followed the basic format and structure of a poem (i.e., was not written in sentence format, was aligned to the center of the page, etc.). The straightforward nature of the item, and the fact that it did not require students to distinguish among various selections, most likely caused this discrepancy in ratings.

Two of the three strands (Comprehending Literary Text and Comprehending Informational Text) consistently rated favorably on the next criterion, range of knowledge correspondence, which evaluates whether a majority of objectives are assessed by the test. The Reading Process strand, however, was rated "weak" on this criterion for all but one grade level (grade 5). The eighth grade assessment provides an example of what causes a "weak" rating on this criterion. While 13 objectives are contained within the Reading Process strand at this grade level, SMEs identified only six objectives assessed by AIMS items.

In addition, in grades 4, 5, 8 and 10, AIMS assessed some objectives multiple times and others only once, as indicated by poor balance of representation ratings. This was most problematic for the Reading Process strand primarily affecting grades 4, 5, and 8. Referring back to SMEs item-objective matches for the eighth-grade test, of the six objectives matched to AIMS items, four were assessed by only one item, one was assessed by two items, and one was assessed by 15 items. Clearly, students whose teachers only address one objective will do well, as long as their teacher selects the right one objective. "Weak" balance of representation ratings were also assigned to Comprehending Information Text at grades 8 and 10 and to Comprehending Literary Text at grade 10.

Finally, the source of challenge issues identified by SMEs were reviewed and their comments summarized for items identified as problematic by at least two SMEs in Table 5. They identified two main concerns: (1) that items did not correspond to any objectives listed in state standards and (2) that an item needed some form of revision such as removing "not" from the stem, correcting a typo, or rewording response options so that only one response is correct. The results of this analysis are presented in Table 5.

For example, reviewers noted that the item, "The story says that Lucy coaxed the dog. Coaxed means about the same as a) remembered, b) patted, c) hugged, d) encouraged," requires students to use context clues to figure out the meaning of "coaxed," a skill not included in the third grade reading standards. In addition, reviewers flagged 3rd grade items 66 and 72 because the items use the word "not" in the stem, a practice they found objectionable. Across all grade levels, SMEs identified 30 items that they could not match to objectives and 27 items that needed some type of revision. All source of challenge comments are included in the binders appending this report.

Table 4

WAT Ratings by Strand and Grade Level, Reading

	Categorical Concurrence			*Depth of Knowledge Consistency*		
	Reading Process	Comprehending Literary Text	Comprehending Informational Text	Reading Process	Comprehending Literary Text	Comprehending Informational Text
Grade 3	Yes	Yes	Yes	Yes	No	Yes
Grade 4	Yes	Yes	Yes	Yes	Yes	Yes
Grade 5	Yes	Yes	Yes	Yes	Yes	Yes
Grade 6	Yes	Yes	Yes	Yes	Weak	Weak
Grade 7	Yes	Yes	Yes	Yes	Yes	Yes
Grade 8	Yes	Yes	Yes	Yes	No	Yes
Grade 10	Yes	Yes	Yes	Yes	Yes	Yes

	Range of Knowledge Correspondence			*Balance of Representation*		
	Reading Process	Comprehending Literary Text	Comprehending Informational Text	Reading Process	Comprehending Literary Text	Comprehending Informational Text
Grade 3	Weak	Yes	Yes	Yes	Yes	Yes
Grade 4	Weak	Yes	Yes	Weak	Yes	Yes
Grade 5	Yes	Yes	Yes	Weak	Yes	Yes
Grade 6	Weak	Yes	Yes	Yes	Yes	Yes
Grade 7	Weak	Yes	Yes	Yes	Yes	Yes
Grade 8	Weak	Yes	Yes	Weak	Yes	Weak
Grade 10	Weak	Yes	Yes	Yes	Weak	Weak

Note: Refer to pages 7-8 to learn how WAT criteria are rated.

Table 5

Source of Challenge Items by Grade Level, Reading

Grade	Item	Does Not Match Objectives	Needs Revision
3	17	X	x
	27	X	
	66		x
	69	X	
	70	X	
	71	X	
	72		x
	79	X	
	80	X	
	90	X	
4	3	X	x
	18	X	
	19	X	
	27	X	x
	28	X	x
	31	X	x
	32		x
	33	X	
	35		x
	36	X	
	62	X	
	68	X	
	81	X	
5	29		x
	33		x
	48		x
	90		x
	91	X	
6	1	X	
	10		x
	35	X	
	49	X	
	72		x
	82	X	x
10	7		x
	10	X	
	14		x
	15		x
	17		x
	18		x
	24		x
	30	X	x
	31	X	x
	35	X	x
	54	X	x
	55	X	x

Mathematics

The categorical concurrence within math strands was generally good, with the exception of Mathematical Structure and Logic. This strand was rated poorly by reviewers across most content alignment measures. For this strand, no categorical concurrence was found at all grade levels except grade 10, indicating that five or fewer items were associated with these skills (Table 6).

Three of the five math strands (Number Sense, Patterns/Algebra/Functions, and Geometry/Measurement) consistently met the depth of knowledge consistency criterion across grade levels, indicating that test items required at least the same level of cognitive challenge as state standards. However, the Mathematical Structure and Logic items on the 3rd through 8th grade AIMS did not achieve depth of knowledge consistency. Grades 3, 6, and 7 were rated "weak" and grades 4, 5, and 8 were rated as having "no" depth of knowledge consistency. Data Analysis and Probability items also lacked the same level of cognitive complexity as their corresponding objectives at half of the grade levels, with this strand rated "weak" at grades 4, 7, and 8. For example, SMEs rated the fourth-grade objective, "develop and algorithm to calculate the perimeter of simple polygons" as requiring strategic thinking, but rated the item they associated with this objective, Item 87, as application of a skill or concept. Item 87 requires the student to choose a response option that could be used to determine the perimeter of a given shape. The student simply needs to recognize the formula to calculate perimeter and plug in the appropriate values from the problem, an application of a skill. The item does not require strategic thinking to develop an algorithm.

AIMS math assessed a majority of objectives (i.e., met the range of knowledge correspondence criterion) for two strands: Data Analysis/Probability and Patterns/algebra/Functions. Approximately half of the grade levels for the other three strands also met this criterion. However, several strand/grade level combinations did not attain range of knowledge correspondence. Grades 4, 5, and 10 had "no" range of knowledge correspondence in Mathematical Structure and Logic and grade 8 was "weak" in this strand. Grades 5 and 10 were also "weak" in range of knowledge correspondence for the Geometry/Measurement strand, as was grade 10 for Data Analysis and Probability and grades 3, 4, and 6 for Number Sense. Grade 7 rated even more poorly in Number Sense, having "no" range of knowledge correspondence. Finally, AIMS math rated positively on the balance of representation measure for each strand at every grade level. This indicates that similar numbers of items were associated with each assessed objective.

As with reading, the math source of challenge issues identified by SMEs were reviewed and their comments summarized for items identified as problematic by at least two SMEs (Table 7). Fewer source of challenge issues were identified in math than were identified in reading when using this approach, although SMEs identified the same two concerns: (1) that items did not correspond to any objectives listed in state standards and (2) that an item needed some form of revision. In mathematics, more items were identified because they did not match objectives (23 items) than because they needed revision (16 items). Common suggestions for revising math items included re-wording the stem to make it clearer, removing sources of bias/culturally insensitive or improving graphics (especially graphs and charts).

Table 6

WAT Ratings by Strand and Grade Level, Mathematics

Categorical Concurrence

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

Depth of Knowledge Consistency

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

Range of Knowledge Correspondence

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

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	Yes
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

For example, Item 14 on the sixth grade test states, “Which of these expressions will have the greatest answer? Use estimation to help you choose the best answer.” Reviewers commented that “the stem was badly written as an expression does not have an answer as currently stated. Also, the item does not match the performance objective it is intended to match.” Reviewers suggested the stem be rewritten to, “use estimation to determine which of the following expressions has the greatest value” to better match the response options and performance objective. Item 83 was identified as a culturally insensitive by reviews. The item uses the content of making a Hopi mask to assess converting within a US measurement system. Reviewers stated that the use of cardboard ice ream tubs and paper towel rolls in making a sacred Hopi mask is disrespectful and felt the item should be omitted.

Table 7

Source of Challenge Items by Grade Level, Mathematics

Grade	Item	Does Not Match Objectives	Needs Revision
3	58	x	x
	70	x	x
4	12	x	
	39		x
	40	x	
	58	x	
	75	x	
	89	x	
6	3	x	
	8	x	
	14	x	x
	30		x
	31	x	x
	32		x
	36	x	x
	41		x
	57	x	
	59		x
	63	x	
	68	x	
	69		x
	70	x	x
	71	x	x
	75	x	
	77	x	x
	83		x
	87	x	
	88	x	
7	40	x	
	54	x	
10	1		x

Special Education

As was shown in Table 3, SMEs tended to assign lower alignment ratings to the AIMS-A special education assessments than to the general AIMS. These lower ratings may reflect the newness of the special education standards, which were in draft form at the time of the review. In fact, most reviewers had not seen the special education standards prior to attending the alignment session. In contrast, SMEs were intimately familiar with the reading and math standards, having implemented them for some time in their classrooms.

In addition, the structure of the special education standards may also account for these lower ratings. To be in accord with the Arizona math and reading standards, the special education standards have separate objectives for each grade from three to eight and high school. However, because most special education students function at a low academic level, a larger number of special education objectives correspond to primary (K-3) grades rather than to higher grade levels. Yet there are only two forms of AIMS-A (Levels I and II)[2] that can be used to assess students from K-12, and consequently, SMEs were given the difficult task of gauging the alignment of items on both forms to objectives that varied from third grade to high school.

The SMEs working with the third grade objectives (the lowest grade level addressed by this study) were given a much larger set of objectives to review than were provided to SMEs at other grade levels. This disproportionate number of objectives might lower the range of knowledge ratings at third grade, since they are based on the percentage of objectives assessed. It is more likely that a test will assess half of the objectives if the standards contain ten objectives (requiring that items match five objectives) than if they contain 100 objectives (requiring that items match 50 objectives). However, we would expect third grade categorical concurrence ratings to be higher due to the larger number of objectives (increasing the probability that at least five items will be associated with that strand). Depth of knowledge consistency and balance of representation ratings should be unaffected by these cross-grade differences.

Special Education Reading

AIMS-A Levels I and II were both rated highly on all WAT criteria for Comprehending Informational Text, in part because the vast majority of items are concerned with this strand (Tables 8 and 9). The other reading strands, reading Process and Comprehending Literary Text, are not assessed as much. Fewer than six items measure Reading Process in grades 4 through 7 as indicated by "no" ratings on categorical concurrence. Comprehending Literary Text is not included in the special education standards except at third grade, and therefore, is not rated using the WAT criterion at other grade levels. Even at third grade, only a small proportion of objectives (13 percent) are concerned with Comprehending Literary Text.

Although they had low categorical concurrence ratings, Reading Process items tended to match the cognitive complexity of objectives, as shown by "yes" depth of knowledge consistency ratings for AIMS-A level 1 and for AIMS-A level 2. Both assessments also received low range of knowledge correspondence ratings for this strand, indicating that they measured only a small

[2] As mentioned earlier in this report, AIMS-A Level I is designed to measure functional and kindergarten standards and AIMS-A Level 2 is designed to measure first through third grade standards.

proportion of objectives. Eighth and tenth grade Reading Process objectives were the exception to this rule; receiving "yes" ratings on AIMS-A Level II and "weak" ratings on AIMS-A Level I. Finally, items were spread equally among those Reading Process objectives that were assessed, garnering "yes" balance of representation ratings.

Table 8

WAT Ratings by Strand and Grade Level, AIMS-A (Special Education) Level I Reading

	Categorical Concurrence			*Depth of Knowledge Consistency*		
	Reading Process	Comprehending Literary Text	Comprehending Informational Text	Reading Process	Comprehending Literary Text	Comprehending Informational Text
Grade 3	Yes	No	Yes	Yes	Yes	Yes
Grade 4	No		Yes	Yes		Yes
Grade 5	No		Yes	Yes		Yes
Grade 6	No		Yes	Yes		Yes
Grade 7	No		Yes	Yes		Yes
Grade 8	Yes		Yes	Yes		Yes
Grade 10	Yes		Yes	Yes		Yes

	Range of Knowledge Correspondence			*Balance of Representation*		
	Reading Process	Comprehending Literary Text	Comprehending Informational Text	Reading Process	Comprehending Literary Text	Comprehending Informational Text
Grade 3	No	No	No	Yes	Yes	Yes
Grade 4	No		Yes	Yes		Yes
Grade 5	No		Yes	Yes		Yes
Grade 6	No		Yes	Weak		Yes
Grade 7	No		Yes	Yes		Yes
Grade 8	Weak		Weak	Yes		Yes
Grade 10	Weak		No	Weak		Yes

Note: Blank cells indicate no objectives exist for that strand at that grade level

The final strand, Comprehending Literary Text (applicable only to grade 3), received consistently low alignment ratings for AIMS-A Level II, with "no" ratings on all WAT criteria. The strand was better aligned to AIMS-A Level I according to SMEs. Although the Level I assessment was categorized as "no" for categorical concurrence and "no" for range of knowledge for the Comprehending Literary Text strand, AIMS-A Level I items were written at the same level of cognitive complexity as this strand's objectives as indicated by a "yes" rating in depth of knowledge consistency. In addition, each Comprehending Literary Text objective assessed by the Level I test was assigned an equal number of items, as indicated by a "yes" rating on balance of representation.

Finally, SMEs identified sources of challenge for a plethora of AIMS-A items on the Level I and Level II assessments. Almost all comments came from reviewers at grades 4 and above and nearly all comments indicated that items did not match objectives (comments are included in the binder appending this report). Other comments were more specific. For example, SMEs

indicated that Items 12 and 19 on grade 4, Level II required students to evaluate written directions; however no objectives dealt with comprehending written directions.

Table 9

WAT Ratings by Strand and Grade Level, AIMS-A (Special Education) Level II Reading

	Categorical Concurrence			*Depth of Knowledge Consistency*		
	Reading Process	Comprehending Literary Text	Comprehending Informational Text	Reading Process	Comprehending Literary Text	Comprehending Informational Text
Grade 3	Yes	No	Yes	Yes	No	Yes
Grade 4	No		Yes	Yes		Yes
Grade 5	No		Yes	Yes		Yes
Grade 6	No		Yes	Yes		Yes
Grade 7	No		Yes	Yes		Yes
Grade 8	Yes		Yes	Yes		Yes
Grade 10	Yes		Yes	Yes		Yes

	Range of Knowledge Correspondence			*Balance of Representation*		
	Reading Process	Comprehending Literary Text	Comprehending Informational Text	Reading Process	Comprehending Literary Text	Comprehending Informational Text
Grade 3	No	No	No	Yes	No	Yes
Grade 4	No		Weak	Yes		Yes
Grade 5	No		Yes	Yes		Weak
Grade 6	No		Yes	Yes		Yes
Grade 7	No		Yes	Yes		Yes
Grade 8	Yes		Yes	Yes		Yes
Grade 10	Yes		Yes	Yes		Yes

Note: Blank cells indicate no objectives exist for that strand at that grade level

Special Education Mathematics

AIMS-A Levels I and II are mainly number sense tests, with high levels of categorical concurrence across grade levels for this strand only. Although the tests also seem to assess some 3rd and 4th grade Geometry/Measurement objectives (Tables 10 and 11). With few exceptions, five or fewer items assess other math strands.

Nonetheless, most items require similar levels of intellectual skill as the objectives they measure, as indicated by "yes" depth of knowledge consistency ratings across most grade levels and strands for AIMS-A Levels I and II. Given the population for whom the tests and objectives were designed, one might expect few higher-order objectives and items. This restriction in range translates to higher depth of knowledge consistency ratings. However, it is important to note that the Level I form of AIMS-A received more "yes" ratings on this criterion than were generated for the Level II test. In addition, the middle school grades (6-8) received most of the "no" ratings on this criterion, indicating one possible area of revision for the draft special education standards.

Both assessments also received low range of knowledge correspondence ratings across strands, indicating that they measured only a small proportion of objectives. This is consistent with the categorical concurrency findings, which showed that few items assessed four of the five math strands. Since most items assess the Number Sense strand, a large proportion of objectives in this strand are expected to be assessed. Closer examination of SMEs' matches of items to objectives revealed an interesting phenomenon. SMEs tended to assign AIMS-A items to Number Sense objectives, but for the fifth through tenth grade objectives, they did not tend to agree on which item was associated with which objective. This may indicate that SMEs believed the items fit somewhere within the Number Sense strand but were not able to find a particularly good objective match for each item. SMEs' source of challenge comments that items did not match objectives (discussed below) is consistent with this interpretation.

Given the poor alignment ratings for the AIMS-A tests, the balance of representation ratings were surprisingly good, with a majority of grade level/strand combinations rated "yes" on this criterion. Two factors are believed to contribute to these ratings. First, balance of representation is only based on those objectives that are assessed by at least one item. Limiting the pool of objectives inflates the balance of representation measure. Second, SMEs tended to assign different items to each objectives (i.e., in an attempt to make an item fit, they matched things up, but did so in ways that were inconsistent with other reviewers). This spread the number of items across a wider range of objectives than would have occurred had item-objective matches been consistent across SMEs.

Finally, as with the reading version of the AIMS-A tests, SMEs indicated that many Level I and Level II items did not match objectives for source of challenge. Comments are included in the binder appending this report. Some source of challenge comments pertained more specifically to substantive issues. For example, three SMEs stated that Item 5 on Level I for third grade required students to match groups with up to ten objects, but the objective only required matching groups with up to five objectives.

Table 10

WAT Ratings by Strand and Grade Level, AIMS-A (Special Education) Level 1 Mathematics

Categorical Concurrence

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

Depth of Knowledge Consistency

	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	No
Grade 5	Yes	Yes	Yes	Yes	Yes
Grade 6	Yes	Yes	Yes	Yes	Weak
Grade 7	Yes	Yes	Yes	Yes	Yes
Grade 8	Yes	Yes	Yes		Yes
Grade 10	Yes	Yes	Yes		Yes

Range of Knowledge Correspondence

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

Balance of Representation

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

Note: Blank cells indicate no objectives exist for that strand at that grade level

Table 11

WAT Ratings by Strand and Grade Level, AIMS-A (Special Education) Level II Mathematics

Categorical Concurrence

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

Depth of Knowledge Consistency

	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	No	Yes
Grade 6	Yes	Yes	No	Yes	No
Grade 7	Yes	Yes	No	No	No
Grade 8	Yes	No	Yes		Yes
Grade 10	Yes	Yes	Yes		Yes

Range of Knowledge Correspondence

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

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	No	No
Grade 6	Yes	Weak	Yes	Yes	No
Grade 7	Yes	Yes	Yes	No	No
Grade 8	Yes	No	Yes		Yes
Grade 10	Yes	Yes	No		Yes

Note: Blank cells indicate no objectives exist for that strand at that grade level

Conclusion and Recommendations

Over the three-day alignment period, SMEs provided valuable information regarding the alignment of the 2005 AIMS and AIMS-A exams to standards. They worked diligently to render accurate and detailed information on the tests and standards. Not only can their feedback be used to judge the 2005 exams, but it can be used to improve future AIMS and AIMS-A exams, and possibly academic standards. It is highly recommended that ADE staff or test company personnel review the source of challenge comments in the appended reports to identify any items that should be dropped or modified. Those items flagged by multiple SMEs should be targeted. 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 subject and grade level, there are some general trends apparent in the tables presented in this report. First and foremost, AIMS was judged to be "aligned" in most grades and across reading and math. In reading (see Table 4) 68 of the possible 84 alignment decisions generated by the WAT model were "Yes," and in mathematics (Table 6), 126 of 140 decisions were considered to be aligned. Overall, AIMS-A also was judged to be aligned by SMEs, but to a lesser extent. Levels I and II for reading were rather promising (68 of 84 "Yes" decisions for Level 1, and 66 of 84 decisions for Level II), but AIMS-A mathematics was more problematic. In Level I math, AIMS-A was judged to be aligned in 62 of 132 decisions, and for Level II math, 64 of the 132 decisions were deemed "Yes."

Certain aspects of AIMS and AIMS-A warrant particular attention for overall test program improvement. In reading, there were three problematic areas. The Comprehending Literacy Text strand in three grades was rated as "weak" or "no," indicating low cognitive demand match between items and objectives. The strand, Reading Process, received "weak" ratings in all but one grade for range of knowledge, and "weak" in three grades for balance of representation. These findings reveal that too few items measure objectives from Reading Process and when items do match to objectives from the strand, a minority of strand objectives are measured by the exams. For future tests, more items should be included that measure Reading Process, and a greater breadth of objectives from that strand should be measured.

In mathematics, Structure and Logic was rated as "no" in categorical concurrence in all but the high school test, indicating a need for greater content match between items and objectives in that area. Also, there were depth of knowledge problems on all but the high school test for this same strand. Items also need to be written with better cognitive demand congruence in Structure and Logic. Depth of knowledge problems also were apparent in three tests for Data Analysis and Probability. Again, there is a need to review the cognitive demand alignment of items and objectives for that strand. Balance of representation was considered sufficient for all tests across all strands, but range of knowledge issues were identified for Number Sense (four tests) and Structure and Logic (four tests).

Keep in mind, however, that one major assumption of the WAT model is that all strands should be measured with more or less the same degree of emphasis. One reason Range of Knowledge might have been an issue for Number Sense and Structure and Logic was the number of items included on the exams to measure those strands. In accord with teachers' wishes, ADE

purposely developed test blueprints that did not reflect equal emphasis across the strands. This decision by ADE must be considered when judging Range of Knowledge alignment with the Webb model.

Taken together, AIMS-A received significantly lower ratings than AIMS exams. This finding is not particular surprising, given that AIMS-A is a newer testing program and has had much less time to develop than AIMS. The information provided here should be reviewed carefully during the revision process of AIMS-A. In reviewing Tables 8 to 11, it is apparent that categorical concurrence and range of knowledge appear to be the two weakest aspects of reading and math AIMS-A Levels I and II. The organization of the special education standards and the fact that only two tests exist to assess the standards across all grade levels, should be taken into account.

The alignment review process identified certain areas that can be addressed to improve AIMS and AIMS-A. It also indicated that Arizona's assessments are working effectively to measure students' attainments of academic standards. In all, especially for AIMS, the tests received high marks on the dimensions of alignment produced through the WAT method. As tests change and improve, alignment studies should become an integral component of the AIMS and AIMS-A revision processes. Feedback from outside experts is necessary to make AIMS and AIMS-A the best testing programs possible for Arizona schools.

Response
to
Alignment Study Findings

Dr. Charles Buren, Dr. Cindy Ziker, Frank Brashear, Judy Crosswell

Arizona Department of Education
Phoenix, Arizona

March, 2006

This is the Arizona Department of Education's response to issues summarized in the Executive Summary of the Alignment study conducted July 27-29, 2005 by Dr. Jerome D'Agostino, using the Webb Alignment Tool. This response addresses issues identified for both the AIMS and AIMS-A.

All quotations are taken from the Alignment Analysis of Arizona Academic Standards and Assessments by Jerome V. D'Agostino, Megan E. Welsh, & Adriana D. Cimetta from the University of Arizona, September 2005.

Source of Challenge

> *"It is highly recommended that ADE staff or test company personnel review the source of challenge comments in the appended reports to identify any items that should be dropped or modified."*

Arizona's Plan:

Regular assessment

Items that were identified as sources of challenge at the high school level were immediately brought to the attention of the item selection committee, and were replaced with appropriately aligned alternate items for the 2006 high school test. This same procedure will be followed for the operational 2007 tests in Grades 3 through 8. Items identified as sources of challenge will have their status changed from "Item Available" to "Do Not Use" in the Item Bank. Item writing by Arizona educators will occur in March for field testing in the spring of 2007. Arizona continues to add items to the Item Bank though annual item writing committee work. "Source of Challenge" items will be included in the training materials for the item writing committees to help illustrate poorly aligned or unclear items. Some NRT items (used for NRT and CRT scores in the 3 through 8 Arizona Instrument to Measure Standards Dual Purpose Assessment -AIMS DPA) were identified as challenge items and will be replaced with aligned NRT items.

Alternate Assessment

The assessment implemented at the time of the alignment study was a transitional assessment to bridge the movement from alternate functional standards to grade level alternate content standards. Implementation of the alternate assessment based on the soon to be approved grade level alternate content standards will occur in 2008. This assessment will address the identified sources of challenge in the alignment study.

Standards Revision

> *"In some cases, source of challenge problems might reveal that objectives need revision."*

Arizona's Plan:

Regular assessment

Arizona will review the academic content standards on a regular five year basis. Since the standards were approved in March of 2003, committees will begin the process of reviewing the standards for mathematics and reading in the summer of 2007 for possible approval in the spring of 2008. Arizona Department of Education -ADE has been collecting feedback on strengths and weaknesses in the current standards since their adoption in 2003. Committees will use the feedback in their deliberations.

Alternate Assessment

Arizona is in the process of revising the current standards for the Alternate Grade Level Academic Content Standards assessed by Arizona's Instrument to Measure Standards Alternate - AIMS-A , both level I and level II. The suggested revisions will link the alternate content standards with the adopted grade level content standards Alternate content standard performance objectives have been created to match the performance objectives from the grade level academic content standards. On March 15, 2006, the alternate assessment committee of educators will review public comments and surveys, make final revisions, and prepare the final document for the State Board. The revisions to these standards will be presented to the State Board in April with approval in May 2006.

Cognitive Demand

> *"In reading, there were three problematic areas. The Comprehending Literacy Text strand in three grades was rated as "weak" or "no," indicating low cognitive demand match between items and objectives."*

Arizona's Plan:

Regular assessment

A deliberate effort was made by ADE to have selection committees choose items for field testing in 2006 that have an increase in the depth of knowledge. These items will be added to the Item Bank in August of 2006.

Arizona commissions all passages for the state assessments. In order to increase the pool of higher level DOK items available for use on the state assessments, Arizona has given guidance to writers of commissioned works to create passages that will allow item writers to create items with greater depth of knowledge. During training of item writers this increase in DOK will be addressed.

Alternate Assessment

The alignment study found that the alternate assessment had a sufficient DOK with the exception of the Level II Comprehending Literary Text Strand. Comprehending Literary Text performance objectives have already been expanded. New items will be created to address these added performance objectives.

> *"Also, there were depth of knowledge problems on all but the high school test for this same strand. Items also need to be written with better cognitive demand congruence in Structure and Logic. Depth of knowledge problems also were apparent in three tests for Data Analysis and Probability. Again, there is a need to review the cognitive demand alignment of items and objectives for that strand."*

Arizona's Plan:

Regular assessment

A gap analysis was performed in December by ADE and the test contractor of items in the current item bank to determine where additional development was needed. As a part of the gap analysis, the DOK of item bank items was determined. Analyzing both the alignment study and the gap analysis results gave ADE a better understanding of what items needed to be produced. As a result, item development committees, composed of Arizona educators, will be instructed to create new items to address the recognized gaps and expand the number of items at specific DOK levels.

Alternate Assessment

In the content area of mathematics, tables identified some weaknesses in the Structure and Logic Strand for Grades 4 and 6, and in Geometry and Measurement for Grades 5 and 7. As part of the task of the standards revision committee, these areas of mathematics were addressed. New items will be created for performance objectives within these strands during the next item writing committee meeting following State Board approval of the standards.

Categorical Concurrence

> *"In mathematics, Structure and Logic was rated as "no" in categorical concurrence in all but the high school test, indicating a need for greater content match between items and objectives in that area."*

Arizona's Plan:

Regular assessment

The blueprint that Arizona had designed for all of the AIMS assessments calls for four items per reported concept. The WAT alignment tool (CD Version 1) required that the number of items per reported strand be six. For Strand 5 in mathematics, Arizona's AIMS for Grades 3-8 contain only 1 concept. Therefore, for strand 5, the number of items per strand is only four and not six. When the default of six was changed to four during the alignment study, no tables were produced for analysis. Since tables were necessary for the study, the default assumption of six needed to be kept. For the alignment study that Dr. D'Agostino conducted, this constraint was imposed. A new version of the WAT alignment tool has been received, and ADE will utilize it in the next alignment study. Due to the awareness of the constraint in the WAT, the blueprint for the Science assessments were revised to meet the six item per strand assumption.

Another alignment study to include Science items and standards, now that the new Science standards have been established, will take place in the Summer of 2007. Items will have been field tested in Spring 2007 and selected for the 2008 operational tests by the beginning of summer.

Alternate Assessment

> *"In reviewing Tables 8 to 11, it is apparent that categorical concurrence and range of knowledge appear to be the two weakest aspects of reading and math AIMS-A Levels I and II."*

Arizona's Plan:

The AIMS –A assessment blueprint mirrors the AIMS blueprint as much as possible, thereby linking the two assessments. For AIMS-A, categorical concurrence weaknesses occurred mostly in the Reading Process in Grades 4 through 7 and Comprehending Literary Text in the same grades. With the new revision of the Alternate Grade Level Academic Content Standards, these areas were addressed. A revised Level I and Level II assessment is planned for Spring 2008. Science will be field tested in 2007, with the operational assessment in place by Spring of 2008.

The portion of the quote concerning Range of knowledge will be addressed below.

Balance of Representation

> *"…and "weak" in three grades for balance of representation. These findings reveal that too few items measure objectives from Reading Process and when items do match to objectives from the strand, a minority of strand objectives are measured by the exams."*

Arizona's Plan:

Regular assessment

Over a period of time, all assessable performance objectives will be tested. The 2005 assessment was used for the alignment study. The 2006 assessment item selection committee deliberately chose items to assess performance objectives that were not represented in the previous assessment. This procedure is followed at all annual item selection committee meetings.

Items from the TerraNova NRT used to generate both NRT and CRT scores, were included in the study. In order to generate a valid TN score the TN blueprint also had to be matched. Having to match two blueprints instead of one affected the perceived emphasis of performance objectives. The available pool of TN items was more limited than the pool of AIMS items and therefore caused more emphasis to be placed on one performance objective over another. Many of these TN items were also sources of challenge. We have entered into discussions with the vendor to rotate out their dual purpose NRT items to improve alignment of the overall test. This replacement of items will affect the balance of representation so that an overemphasis of one concept over another will not occur.

All AIMS items were recently recoded to the adopted standards. Many items changed after the recoding meetings. Since the alignment study preceded the recoding committee meetings, items that were perceived as overemphasizing one performance objective as well as being miscoded, were considered sources of challenge. Recoding has resolved this issue. These challenged items will be rotated out in subsequent assessments since they no longer match the test blueprint.

Alternate assessment

The alternate assessment, AIMS-A, in its balance of representation, parallels the regular AIMS assessments. Since a committee has revised and expanded the performance objectives, additional items will be incorporated into the test for the 2008 assessment. This will change the emphasis and, in turn, positively affect the balance of representation.

Range of Knowledge

> *"…, but range of knowledge issues were identified for Number Sense (four tests) and Structure and Logic (four tests)."*

Arizona's Plan:

Regular assessment

Further examination of the Item Agreement Coverage table for some grades yielded the observation that some reviewers, by virtue of the fact that they could choose multiple performance objectives per assessment item, chose across strand objectives. This affected the range of knowledge by not having the item correctly placed in the single strand it was coded for. Since then, the items have been recoded, this discrepancy should not occur again in the next alignment study.

All of the performance objectives are expected to be assessed on the local level; however, at the state level, not all of the objectives are appropriate to be assessed. When the blueprints were established, a committee of Arizona educators determined which performance objectives were to be assessed, by the statewide assessment, and in what order. The WAT assumes that at least fifty percent of all objectives within a strand will be assessed. Since the WAT does not make this distinction of local versus state, nor does it make any allowance for rotating performance objectives assessed, the number of performance objectives identified is larger than the number of performance objectives assessed on a statewide basis in a single year. The WAT reporting of the range of knowledge is definitely adversely affected by the Arizona assessment plan.

The researchers supported Arizona's decision regarding the blueprints by stating the following:

> *"Keep in mind, however, that one major assumption of the WAT model is that all strands should be measured with more or less the same degree of emphasis. One reason Range of Knowledge might have been an issue for Number Sense and Structure and Logic was the number of items included on the exams to measure those strands. In accord with teachers' wishes, ADE purposely developed test blueprints that did not reflect equal emphasis across the strands. This decision by ADE must be considered when judging Range of Knowledge alignment with the Webb model."*

Alternate assessment

> *"For future tests, more items should be included that measure Reading Process, and a greater breadth of objectives from that strand should be measured."*
> *"In reviewing Tables 8 to 11, it is apparent that categorical concurrence and range of knowledge appear to be the two weakest aspects of reading and math AIMS-A Levels I and II."*

To address the researchers' comments for breadth of objectives, item writers for the alternate assessment will be pooled with item writers for the regular assessment to assist in creating items that cover a wider range and depth of knowledge. Special education teachers will also be paired with AIMS item writers as mentors to write new items for AIMS-A. ADE feels that this process

will create a closer match to the grade level AIMS assessment. By doing this, cross training and development of a cadre of highly qualified item writers will benefit both the regular as well as the alternate assessments.

The expansion of standards necessitates more item writing and better item writing skills.

Test length was a major consideration in the establishment of the test blueprint. While the number of performance objectives was not limited, the number of test items was limited. When the blueprints were established for the current assessments, the NCTM weighting chart was considered along with the Arizona Academic Content Standard for Mathematics. This weighting limited the number of performance objectives assessed in each strand. The WAT alignment tool assumes that all strands have equal representation. This is not really true of mathematics.