Alignment of Mathematics State-level Standards and Assessments: The Role of Reviewer Agreement

CSE Report 685

Noreen Webb and Joan Herman
University of California

Norman Webb
University of Wisconsin, Madison

June 2006

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE),
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.3 Indicators of Classroom Practice and Alignment, Strand 1: Methodological Issues/Directors: Herman, Wang, & Wells.

Copyright © 2006 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

Alignment of Mathematics State-level Standards and Assessments: The Role of Reviewer Agreement

Noreen Webb and Joan Herman
University of California

Norman Webb
University of Wisconsin, Madison

Abstract

In this report we explore the role of reviewer agreement in judgments about alignment between tests and standards. Specifically, we consider approaches to describing alignment that incorporate reviewer agreement information in different ways. The essential questions were whether and how taking into account reviewer agreement changes the picture of alignment between tests and standards. This study showed a wide range of reviewer agreement during the process of aligning standards and assessments, with substantial reviewer disagreement about important elements such as correspondence between objectives and items on the assessments. Taking this reviewer disagreement into account changed conclusions about alignment, not only showing weaker alignment than previously demonstrated, but also changing the profiles of alignment about, for example, relative coverage of specific standards. The results of this study point to the need for greater clarity in objectives and standards, more extensive reviewer training during the alignment process, and possibly also inspection of items to uncover characteristics that may lead to uncertainty among reviewers.

The alignment of standards and assessment is key to today's standards-based reform where assessment serves as both a lever and a measure for the reform effort. State assessments send strong signals to schools about what they should be teaching and what students should be learning, and schools respond by teaching what is assessed (Herman, 2004; Koretz, Barron et al., 1996; Koretz, Mitchell et al., 1996; McDonnell & Choissier, 1997; Lane et al., 2000; Stecher et al. 2000). At the same time, assessment results are expected to provide accurate information to the public, its policy makers, educators, parents and students themselves about how students are doing, and to provide stakeholders with important feedback on which to base their improvement efforts. Absent strong alignment between standards and assessments, schools may ignore desired standards and instead teach only what is tested. Moreover, if what is tested does not well reflect expectations for student performance, test results cannot

provide accurate data about students' or schools' progress relative to those expectations, and improvement actions based on such results are unlikely to further intended goals. Recognizing these key validity concerns, federal Title I legislation since 1994 has required the alignment of standards and state assessments, and current regulations under No Child Left Behind require states to conduct alignment studies to document the technical quality of their tests.

Pioneered by Andrew Porter and Norman L. Webb, systematic procedures for assessing alignment have been well developed (Ananda, 2003; Bhola, Impara, & Buckendahl, 2003; Herman, N. M. Webb, & Zuniga, 2003, 2005; Olson, 2003; Porter & Smithson, 2001; Rothman et al., 2002; N. L. Webb, 1997, 2002, 2005) and now are being applied in states across the country. In essence, these approaches convene panels of experts to analyze assessment items against a matrix defined by an exhaustive set of topics comprising a subject area domain and by levels of cognitive demand, reflecting a range from rote memory to procedures, applications, and complex problem-solving. The matrices then become the basis for computing various indices of alignment to convey how well a test reflects intended standards. Yet, while the process rests firmly on expert or reviewer judgment, basic questions about the reliability of the process have not yet been fully addressed. In particular, the extent to which disagreements among experts, as well as among stakeholder groups carrying out alignment ratings, may influence alignment conclusions have remained unexamined (Buckendahl, Plake, Impara, & Irwin, 2000; Porter, 2002; Porter & Smithson, 2001; N. L. Webb, 1997, 1999, 2002).

This study, then, explores the role of reviewer agreement in judgments about alignment between tests and standards. Specifically, we explore approaches to describing alignment that incorporate reviewer agreement information in different ways. The essential questions were whether and how taking into account reviewer agreement changes the picture of alignment between tests and standards.

Method

Alignment Criteria and Approaches for Assessing Alignment

Approach 1. This study investigated multiple ways to form conclusions about alignment between tests and standards. The first analytic approach was that used by N.

L. Webb (1997, 1999, 2002, 2005) in which reviewers coded one depth-of-knowledge level to each assessment item, and identified each assessment item as corresponding to up to three objectives. The averages of the reviewers' ratings were used to determine whether the alignment criteria were met. All reviewers' ratings entered analyses of the alignment between standards and assessments on the following criteria.

Categorical concurrence between standards and assessment refers to the same or consistent categories of content appearing in both standards and an assessment. Categorical concurrence for a standard was determined using the mean number of items reviewers coded as corresponding to objectives under that standard. An acceptable level of categorical concurrence between a standard and an assessment was declared if reviewers, on the average, rated at least six items as measuring content from the standard (see Webb, 2005, for a justification of this number).

Depth-of-knowledge consistency between standards and assessment refers to a match between the cognitive demands of the standards and an assessment. Depth-of-knowledge consistency for a standard was declared if at least 50% of test items corresponding to a standard were at or above the knowledge level of the objectives which reviewers coded as corresponding to those items. The levels of depth of knowledge that reviewers assigned are as follows (from Webb, 2005).

Level 1 (Recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify a Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels, depending on what is to be described and explained.

Level 2 (Skill/Concept) includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects.

Some action verbs, such as "explain," "describe," or "interpret," could be classified at different levels, depending on the object of the action. For example, interpreting information from a simple graph, or requiring the reading of information from the graph, also are at Level 2. Interpreting information from a complex graph that requires some decisions on which features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited only to number skills, but can involve visualization skills and probability skills. Other Level 2 activities include noticing and describing non-trivial patterns; explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

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

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

between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Range-of-knowledge for a standard refers to the span of knowledge that students need in order to correctly answer the assessment items. Range of knowledge for a standard was determined using the number of objectives that reviewers coded for that standard. Range-of-knowledge correspondence was declared if at least 50% of objectives for a standard, on the average, had at least one related assessment item. That is, if the mean number of objectives that reviewers rated as being measured by items on the test was at least 50% of the objectives for that standard, range of knowledge correspondence was declared as being met . If between 40% to 50% of the objectives for a standard have a corresponding assessment item, the criterion is "weakly" met (for a justification of these levels, see Webb, 2005).

Approach 2. The second alignment approach applied the same alignment criteria as described above, but used only sets of ratings for which item-objective (or item-standard) correspondences met a minimum level of reviewer agreement. First, to arrive at judgments about *categorical concurrence*, reviewer agreement about the specific items matched to each objective and reviewer agreement about the specific items matched to each standard were both taken into account. To take into account agreement about specific item-objective matches, only items for which a minimum number of reviewers agreed on the objective matched to each item were included. Two thresholds of reviewer agreement were used: a bare majority (5 of 9 reviewers in Dataset 1, 11 of 20 reviewers in Dataset 2) and a clear majority (6 of 9 reviewers in Dataset 1; 13 of 20 reviewers in Dataset 2). For example, for bare-majority agreement in dataset 1, only items were considered for which at least 5 reviewers agreed on the objective corresponding to each item. The score for a reviewer used in the analysis was the number of items for which the reviewer agreed with the majority of reviewers on the objective assigned to each item. The mean number of items, averaged over all reviewers, was used to declare categorical concurrence.

Similarly, to take into account agreement about specific item-*standard* matches, only items for which a minimum number of reviewers agreed on the *standard* matched to each item were included in the analysis. The score for a reviewer was the number of items for which the reviewer agreed with the majority of reviewers on the standard assigned to each item. The mean number of items, averaged over all reviewers, was used to declare categorical concurrence.

For depth of knowledge consistency, only items for which a minimum number of reviewers agreed on the item-objective match were included in the analysis. Depth-of-knowledge consistency for a standard was declared if at least 50% of test items corresponding to a standard were at or above the knowledge level of the objectives reviewers coded as corresponding to those items.

For range-of-knowledge correspondence, only items for which a minimum number of reviewers agreed on the specific correspondence between item and objective were included in the analysis. The rating for a reviewer was the number of objectives for which the reviewer agreed with the majority of reviewers on the objective assigned to an item. The mean number of objectives, averaged over all reviewers, was used to declare range-of-knowledge correspondence. If the mean number of objectives exceeded 50% of the total number of objectives for a standard (e.g., geometry and measurement), alignment was judged to be acceptable.

Data Sources

The data used in the analyses come from two recent alignment studies. The first study (N. L. Webb, 2005) used nine reviewers to evaluate the alignment between Michigan's high school mathematics standards and six different assessments; the Michigan Educational Assessment Program high school test is analyzed here. Reviewers participated in a consensus process to determine the depth-of-knowledge levels of the Michigan high school objectives, and then individually matched assessment items to objectives, goals, and standards, and identified depth of knowledge of assessment items. The Michigan high school mathematics standards list six standards (e.g., patterns, relationships, and functions) with up to 18 objectives (e.g., analyze and generalize mathematical patterns including sequences, series and recursive patterns) for each one.

The second set of data analyzed here (Herman et al., 2005) came from a study of alignment between the Golden State Examination (GSE) in high school mathematics and the University of California Statement of Competencies in Mathematics (competencies expected for entering freshmen). Twenty reviewers individually rated the mathematics items of the GSE relative to the expectations identified in the UC competency statement, identifying item features related to content and depth of knowledge (as well as an additional item feature, centrality). The University of California Statement of Competencies in Mathematics lists six content categories (e.g., variables, equations, and algebraic expressions) with up to 10 specific topics considered

essential for entering college freshmen (e.g., solutions of linear equations and inequalities) in each content category.

Results

Alignment between Michigan High School Mathematics Standards and the Michigan Educational Assessment Program (MEAP) (Dataset 1)

Categorical concurrence. Table 1 provides information about categorical concurrence, the number of test items that correspond to a standard. The first columns of Table 1 do not require a minimum level of reviewer agreement about the match between an item and an objective or standard. The mean number of items is the mean, over reviewers, of the number of items corresponding to any objective under a standard.[1] For example, reviewers found that 9.78 items, on average, corresponded to objectives under Standard I. Using this approach, reviewers were not required to agree on the objective measured by a particular item. If a reviewer judged that a particular item measured an objective under Standard I, it did not matter whether other reviewers made a different judgment about the objective measured by that item, or even a different judgment about the standard measured by that item. Using this approach shows categorical concurrence for five of six standards. That is, five of six standards were judged to be measured by at least six items. It should be noted that the total number of items exceeds the number of items on the test (43) because reviewers could match an item to objectives in more than one standard.

[1]These results differ slightly from those reported by N. L. Webb (2005). In that report, a reviewer assigning an item to two objectives from the same standard received a score of 2 (the item was counted twice for the same standard). In the current analyses, the reviewer received a score of 1 (the item was counted once for a standard). Compare, for example, 9.78 in Table 1 to 10.44 reported in N. L. Webb (2005).

Table 1

Categorical Concurrence: Michigan MEAP Mathematics/ Taking into Account Reviewer Agreement on Item-*Objective* Match

	Level of Reviewer Agreement Required for Item-Objective Match								
	None			5 of 9 Reviewers (56% Agreement)			6 of 9 Reviewers (67% Agreement)		
Standard	# items		Categorical Concurrence[2]	# items		Categorical Concurrence	# items		Categorical Concurrence
	M[1]	*SD*		*M*	*SD*		*M*	*SD*	
I. Patterns, relationships and functions	9.78	2.17	YES	2.67	1.11	NO	1.56	.73	NO
II. Geometry and measurement	8.33	1.22	YES	2.89	.78	NO	2.33	.71	NO
III. Data analysis and statistics	8.89	1.17	YES	4.33	1.50	NO	3.33	.87	NO
IV. Number sense and numeration	2.67	1.66	NO	1.33	.50	NO	1.33	.50	NO
V. Numerical and algebraic operations and analytical…	7.56	2.24	YES	2.00	1.00	NO	.89	..33	NO
VI. Probability and discrete mathematics	6.89	1.26	YES	3.78	1.20	NO	2.67	.71	NO

[1] The number of items in this column totals more than the number of items on the assessment (43) because reviewers could select up to three objectives for each item.

[2]YES = mean number of items is 6 or greater; NO = mean number of items is less than 6.

The results are considerably different when a minimum level of reviewer agreement is required. The analyses generating the second set of columns in Table 1 require at least 5 of 9 reviewers to agree on the match between an item and objective. For example, reviewers found that 2.67 items, on average, corresponded to objectives under Standard I *and* they agreed on the specific objectives measured by those items—this is many fewer than the 9.78 items using the previous approach. These results show that, while reviewers, on average, judged that about 10 items measured objectives under Standard I, they disagreed about the objectives measured by most of them. Requiring a bare majority of reviewers to agree on the objective measured by an item, then, vastly reduced the number of items corresponding to each standard. Requiring a greater majority of reviewers to agree on the objective measured by an item (6 of 9 reviewers) reduced the number of items still further (1.56 in Table 1). In both cases, categorical concurrence was not met for any standard. Requiring reviewers to agree on the match between item and objective, then, produced a picture of few items measuring each standard.

It can be argued that requiring agreement on the objective tied to an item is too strict a yardstick for categorical concurrence because the goal is to determine the number of items that correspond to each *standard.* That is, if reviewers agree that an item measures some objective under a particular standard (even if they don't agree on the particular objective), that result will suffice for categorical concurrence. Table 2, therefore, presents information about categorical concurrence when reviewers are required to agree only on whether an item measures some objective under a standard, even if they disagree about the particular objective.

Table 2

Categorical Concurrence: Michigan MEAP Mathematics/ Taking into Account Reviewer Agreement on Item-*Standard* Match

Standard	Level of Reviewer Agreement Required for Item-Standard Match								
	None			5 of 9 Reviewers (56% Agreement)			6 of 9 Reviewers (67% Agreement)		
	# items		Categorical Concurrence	# items		Categorical Concurrence	# items		Categorical Concurrence
	M	SD		M	SD		M	SD	
I. Patterns, relationships and functions	9.78	2.17	YES	7.11	.78	YES	6.00	.00	YES
II. Geometry and measurement	8.33	1.22	YES	7.56	.53	YES	7.00	.00	YES
III. Data analysis and statistics	8.89	1.17	YES	7.22	.97	YES	6.67	.71	YES
IV. Number sense and numeration	2.78	1.79	NO	1.44	.53	NO	1.44	.53	NO
V. Numerical and algebraic operations and analytical…	7.56	2.24	YES	5.11	1.62	NO	5.11	1.62	NO
VI. Probability and discrete mathematics	6.89	1.27	YES	5.89	.33	NO	5.89	.33	NO

The results in Table 2 show that requiring reviewer agreement produced a picture of categorical concurrence for half of the standards, as opposed to 5 of 6 standards when reviewer agreement is not required. The results also show that, for substantial numbers of items, reviewers did not agree on the standard that was measured by the item. For example, reviewers, on average, found that 9.78 items measured objectives under Standard I (patterns, relationships, and functions) when they were not held to any agreement yardstick. When reviewers were required to agree that an item measured some objective under Standard I (even if they didn't agree on the particular objective), that number fell to 7.11 for an agreement level of 5 of 9 reviewers, and fell to 6.00 for an agreement level of 6 of 9 reviewers. These results show that reviewers disagreed about a substantial number of items that measured Standard I. In fact, some items were classified as matching objectives under as many as four standards across the 9 reviewers.

Depth-of-knowledge consistency. Table 3 presents the results for depth-of-knowledge consistency, the degree to which depth of knowledge of the test items meet or exceed the depth of knowledge of the objectives. The first set of columns in Table 3 give the results concerning depth-of-knowledge consistency when agreement between reviewers about the objective measured by the item is not required. The assessment shows depth-of-knowledge consistency for 4 out of 6 standards. For those standards, at least 50 percent of test items for a standard were judged to be at or above the depth of knowledge assigned to the corresponding objectives.

Requiring reviewer agreement about the specific objectives measured by the items produced a similar picture of depth-of-knowledge consistency (Table 3). For four of six standards, at least 50 percent of test items were judged to require depth of knowledge at or above the depth of knowledge assigned to the corresponding objectives.

Table 3

Depth of Knowledge Consistency Between Standards and Assessment: Michigan MEAP Mathematics

Standard	Level of Reviewer Agreement Required														
	No Reviewer Agreement Required					5 of 9 Reviewers (56% Agreement)					6 of 9 Reviewers (67% Agreement)				
	# of items		% Items At/Above DOK of Objective		DOK Cons. [1]	# of items		% Items At/Above DOK of Objective		DOK Cons.	# of items		% Items At/Above DOK of Objective		DOK Cons.
	M	SD	M	SD		M	SD	M	SD		M	SD	M	SD	
I. Patterns, relationships and functions	9.78	2.17	9	11	NO	2.67	1.11	0	0	NO	1.56	.73	0	0	NO
II. Geometry and measurement	8.33	1.22	81	8	YES	2.89	.78	100	0	YES	2.33	.71	100	0	YES
III. Data analysis and statistics	8.89	1.17	31	13	NO	4.33	1.50	37	22	NO	3.33	.87	46	29	WEAK
IV. Number sense and numeration	2.78	1.79	73	34	YES	1.33	.50	72	44	YES	1.33	.50	72	44	YES
V. Numerical and algebraic operations and analytical…	7.56	2.24	69	18	YES	2.00	1.00	79	31	YES	.89	.33	88	35	YES
VI. Probability and discrete mathematics	6.89	1.27	62	25	YES	3.78	1.20	67	36	YES	2.67	.71	67	41	YES

[1] YES = at least 50% of items have DOK assignments at or above DOK of corresponding objectives.

WEAK = between 40% and 50% of items have DOK assignments at or above DOK of corresponding objectives.

NO = less than 40% of items have DOK assignments at or above DOK of corresponding objectives.

However, it is important to note that these conclusions about depth of knowledge are based only on a minority of items on the test: 17 items out of a total of 43 items (40%), averaged over reviewers using an agreement level of 5 of 9 reviewers, and about 12 items (28%) using an agreement level of 6 of 9 reviewers. The majority of test items did not enter the depth-of-knowledge analyses reported in the latter columns of Table 3 because reviewers did not sufficiently agree on the objective corresponding to the item. So, even though the conclusions about depth of knowledge for each standard did not change much when reviewer agreement was required, the depth of knowledge of the majority of items on the test could not be determined when rater agreement was required because too few reviewers agreed on the item that measured an objective.

Range-of-knowledge correspondence. Table 4 presents information about range-of-knowledge correspondence, the percent of objectives in a standard that are represented on the test. The first set of columns in Table 4 present the range-of-knowledge results when reviewer agreement is not required. These results show that only one standard (Standard V) met the yardstick for acceptable range-of-knowledge correspondence, at least 50% of the objectives under a standard being represented on the test. For the remaining standards, fewer than half of the objectives were represented on the test.

Requiring a minimum level of reviewer agreement about the objective matching an item produced a similar pattern: none of the standards met the yardstick for range-of-knowledge correspondence. Moreover, the percent of a standard's objectives that were represented on the test dropped dramatically (Table 4).

Table 4

Range-of-Knowledge Correspondence: Michigan MEAP Mathematics

		Level of Reviewer Agreement Required for Item-Objective Match														
		None					5 of 9 Reviewers (56% Agreement)					6 of 9 Reviewers (67% Agreement)				
		# Obj.		% of Total			# Obj.		% of Total			# Obj.		% of Total		
Standard	# Objs	M	SD	M	SD	Range of Know.[1]	M	SD	M	SD	Range of Know.	M	SD	M	SD	Range of Know.
I. Patterns, relationships and functions	11	4.22	1.20	38	11	NO	1.44	.73	13	7	NO	.89[2]	.33	9	3	NO
II. Geometry and measurement	18	5.78	.67	32	4	NO	2.89	.78	16	4	NO	2.33	.71	13	4	NO
III. Data analysis and statistics	14	5.00	1.12	36	8	NO	3.44	.73	25	5	NO	3.33	.87	24	6	NO
IV. Number sense and numeration	14	2.44	1.42	17	10	NO	1.33	.50	10	4	NO	1.33	.50	10	4	NO
V. Numerical and algebraic operations and analytical…	9	5.22	1.56	58	7	YES	2.00	1.00	22	11	NO	.89	.33	10	4	NO
VI. Probability and discrete mathematics	11	3.67	1.22	33	11	NO	1.78	.44	16	4	NO	1.00	.00	9	0	NO

[1]YES = at least 50% of objectives for a standard had at least one related assessment item.
NO = fewer than 50% of objectives for a standard had at least one related assessment item.

Alignment between the *University of California Statement of Competencies in Mathematics* and the Golden State Examination (GSE) (Dataset 2)

Categorical concurrence. Table 5 provides information about categorical concurrence for the GSE, the number of test items that correspond to a content category on the *University of California Statement of Competencies in Mathematics*. The first columns of Table 5 do not require a minimum level of reviewer agreement about the match between an item and a topic or content category. The mean number of items is the mean, over reviewers, of the number of items corresponding to any topic in a content category. For example, reviewers found that 15.70 items, on average, corresponded to topics in Content Category I. When reviewers were not required to agree on the topic (or content category) measured by a particular item, 3 of 6 content categories showed categorical concurrence. That is, 3 of 6 content categories were judged to be measured by at least six items. It should be noted that the total number of items exceeds the number of items on the test (42) because reviewers could match an item to topics in more than one content category.

The results are considerably different when a minimum level of reviewer agreement is required. The analyses generating the second set of columns in Table 5 require at least 11 of 20 reviewers to agree on the match between an item and topic. For example, reviewers found that 6.25 items, on average, corresponded to topics in Content Category I *and* they agreed on the specific topics measured by those items. This is many fewer than the 15.70 items using the previous approach. These results show that, while reviewers, on average, judged that about 15 items measured topics in Content Category I, they disagreed about the specific topics measured by most of them. Requiring a bare majority of reviewers to agree on the objective measured by an item, then, vastly reduced the number of items corresponding to each standard. Requiring a greater majority of reviewers to agree on the objective measured by an item (13 of 20 reviewers) reduced the number of items still further (5.90 in Table 5). The more stringent reviewer agreement yardstick reduced the number of content categories for which categorical concurrence was met: from three to two.

Table 5

Categorical Concurrence : Golden State Exam in Mathematics/ Taking into Account Reviewer Agreement on Item-*Topic* Match

Content Category	Level of Reviewer Agreement Required for Item-Topic Match								
	None			11 of 20 Reviewers (55% Agreement)			13 of 20 Reviewers (65% Agreement)		
	# items[1]		Categorical Concurrence[2]	# items		Categorical Concurrence	# items		Categorical Concurrence
	M	*SD*		*M*	*SD*		*M*	*SD*	
I. Variables, equations, and algebraic expressions	15.70	4.03	YES	6.25	.97	YES	5.90	.79	NO
II. Families of functions and their graphs	10.85	1.87	YES	7.95	1.50	YES	7.35	1.27	YES
III. Geometric concepts	11.25	1.59	YES	9.75	1.71	YES	8.55	1.47	YES
IV. Probability	5.70	1.26	NO	2.85	.49	NO	2.85	.49	NO
V. Data analysis and statistics	1.90	1.44	NO	.75	.44	NO	.75	.44	NO
VI. Argumentation and proof	.30	.73	NO	.00	.00	NO	.00	.00	NO

[1] The number of items in this column totals more than the number of items on the assessment because reviewers could select up to two topics for each item.

[2]YES = mean number of items is 6 or greater; NO = mean number of items is less than 6.

As was the case for the Michigan results (Dataset 1), it can be argued that requiring agreement on the specific item tied to an item is too strict a yardstick for categorical concurrence because the goal is to determine the number of items that correspond to each *content category.* That is, if reviewers agree that an item measures some topic in a particular content category (even if they don't agree on the particular topic), that will suffice for categorical concurrence. Table 6, therefore, presents information about categorical concurrence when reviewers are required to agree only on whether an item measures some topic in a content category, even if they disagree about the particular topic that the item measures. The results of Table 6 show that, despite the reduction in the number of items found to match a content category when a rater agreement yardstick is applied, the overall picture of categorical concurrence remains the same: half of the content categories exceed the threshold for categorical concurrence; at least six items measure topics in a content category.

Table 6

Categorical Concurrence: Golden State Exam in Mathematics/ Taking into Account Reviewer Agreement on Item-*Content* Category Match

Content Category	Level of Reviewer Agreement Required for Item-Content Category Match: None			Level of Reviewer Agreement Required for Item-Content Category Match: 11 of 20 Reviewers (55% Agreement) and 13 of 20 Reviewers (65% Agreement)[1]		
	# items		Categorical Concurrence[2]	# items		Categorical Concurrence
	M	SD		M	SD	
I. Variables, equations, and algebraic expressions	15.70	4.03	YES	8.40	1.23	YES
II. Families of functions and their graphs	10.85	1.87	YES	9.10	1.29	YES
III. Geometric concepts	11.25	1.59	YES	9.85	1.14	YES
IV. Probability	5.70	1.26	NO	4.80	.89	NO
V. Data analysis and statistics	1.90	1.44	NO	.75	.44	NO
VI. Argumentation and proof	.30	.73	NO	.00	.00	NO

[1]The results are the same for both levels of reviewer agreement.

[2]YES = mean number of items is 6 or greater; NO = mean number of items is less than 6.

Range-of-knowledge correspondence. Table 7 presents information about range-of-knowledge correspondence, the percent of topics in a content category that are represented on the test. The first set of columns in Table 7 present the range-of-knowledge results when reviewer agreement is not required. These results show that four content categories met the yardstick for acceptable range-of-knowledge correspondence, at least 50% of the objectives under a standard being represented on the test. For the remaining two content categories, fewer than half of the topics were represented on the test.

Requiring a minimum level of reviewer agreement about the topic matching an item produced a much different pattern of results. For both reviewer agreement yardsticks (agreement among 55% reviewers, agreement among 65% of reviewers), *none* of the content categories met the yardstick for range-of-knowledge correspondence (Table 7). That is, when reviewers were required to agree on the topic measured by an item, the percentage of topics represented on the test was lower than half for every content category.

Table 7

Range-of-Knowledge Correspondence: Golden State Exam in Mathematics

		Level of Reviewer Agreement Required for Item-Topic Match														
		None					11 of 20 Reviewers (55% Agreement)					13 of 20 Reviewers (65% Agreement)				
		# Topics		% of Total			# Topics		% of Total			# Topics		% of Total		
Standard	# Topics	M	SD	M	SD	Range of Know.[1]	M	SD	M	SD	Range of Know.	M	SD	M	SD	Range of Know.
I. Variables, equations, and algebraic expressions	8	5.75	1.45	72	18	YES	3.35	.59	42	7	NO	2.80	.41	35	5	NO
II. Families of functions and their graphs	10	5.45	1.28	55	13	YES	3.60	.60	36	6	NO	3.60	.60	36	6	NO
III. Geometric concepts	10	5.45	1.00	55	10	YES	4.40	.68	44	7	NO	3.80	.41	38	4	NO
IV. Probability	5	2.95	.69	59	14	YES	2.00	.00	40	0	NO	1.95	.22	39	4	NO
V. Data analysis and statistics	4	1.40	.94	35	24	NO	.75	.44	19	11	NO	.75	.44	19	11	NO
VI. Argumentation and proof	4	.30	.73	8	18	NO	.00	.00	00	0	NO	.00	.00	0	0	NO

[1]YES = at least 50% of objectives for a standard had at least one related assessment item.

NO = fewer than 50% of objectives for a standard had at least one related assessment item.

Discussion

Our results showed that the two approaches to analyzing alignment (based on reviewer means using all ratings on all items vs. requiring a minimum threshold of reviewer agreement) yielded somewhat different judgments about alignment. Effects were seen for both categorical concurrence, the number of items on the test corresponding to each standard (or content category) and range of knowledge correspondence, the percentage of objectives (or topics) represented on the test. Requiring reviewers to agree about the objective (or topic) measured by an item reduced the pictures of both categorical concurrence and range of knowledge; that is, producing a picture of reduced alignment. Changing the exact level of reviewer agreement—in this case a bare majority (slightly larger than 50%) vs. a greater majority (about two-thirds)—had little effect on the results.

The results point to the large disagreement among reviewers about the particular objective(s) that an item measured. On the MEAP, on only 15 of the 43 (35%) items did two-thirds of the reviewers agree on the specific objective measured. On the GSE, two-thirds of reviewers agreed on the specific topic measured by 30 (71%) of the items. One reason may be the ambiguity or lack of clarity of the wording of the objectives. For example, some items on the MEAP were matched to the objective "analyze and generalize mathematical patterns including sequences, series and recursive patterns" by some reviewers, and to the objective "use patterns and reasoning to solve problems and explore new content" by other reviewers. An item that involved generalizing or reasoning from a mathematical series might be classified variously depending on a reviewer's interpretation of the phrase "explore new content". While agreement among reviewers about the standard (or content category) represented by an item was higher, there still was substantial disagreement, especially on the MEAP where two-thirds of reviewers disagreed on the standard corresponding to 7 (16%) of the items.

A philosophical issue is whether reviewer agreement should be taken into account when creating pictures of alignment. If we are trying to describe how many items on a test measure a particular standard (categorical concurrence), it is important that reviewers reach at least some minimum level of agreement about whether each item measures some objective corresponding to that standard. Consider, for example, an item for which half of the reviewers see it as measuring Standard A (but not Standard B), and half of the reviewers classify it as measuring Standard B (but not

Standard A). Matching that item to Standard A for some reviewers and to Standard B for other reviewers is a concern that deserves further investigation.

If we are trying to determine how many objectives in each standard are represented on the test (range of knowledge), it is important that reviewers reach at least a minimum level of agreement about the objective tested by an item. As an extreme case, if some reviewers judged that six objectives in a standard were represented on the test and other reviewers judged that a completely different set of objectives in a standard were represented on the test, we would have little confidence in which objectives were actually represented, and the validity of the test as well as the alignment process would be at issue.

In the case of such discrepancies between reviewers, further exploration is necessary. What are the sources of the inconsistencies? Do reviewers need additional training? Do the characteristics of items contribute to uncertainty among reviewers about the particular objective(s) being measured? Is it possible that some problems actually can be solved in multiple ways? For example, in our study of the Golden State Examination, there were items that could be solved either through algebraic or geometric reasoning, and depending on raters' penchants, were matched to either one or the other standard.

More likely, the clarity and precision of the standards and objectives themselves are a large part of the problem. Standards and objectives, as they are worded, can be vague and/or ambiguous, leaving alignment raters with a Rorschach in projecting intended meaning and thus match with specific items. Worse yet, teachers are faced with the same Rorschach in attempting to teach to the standards. Little wonder, then, that they may turn to teaching to the specifics of the test, rather than the standards. As the National Research Council's Committee on Test Design for K-12 Science Assessment (Wilson & Berenthal, 2005) noted:

> The one general principle that emerged from the committee's review of state ….standards is the need for clear, thorough, understandable descriptions. For standards to play a central role in assessment and accountability systems, they must communicate clearly to all stakeholders in the system – teachers, assessment developers, students, parents and test developers what students are expected to know and be able to do (p.58).

The committee went on to advocate for elaborated standards that were clear, detailed and complete in specifying content and performance expectations, as well as feasible

and grounded in a conceptual framework that was based on student cognition and learning.

Such clarification of standards likely would not only improve the reliability of the alignment process but also has the potential to improve the alignment of state assessments with standards. Test developers, who may currently have the same varied interpretations of what is expected as did the raters in this study, would have a clearer blue print against which to develop test specifications and items. The standards could guide initial test development and strong alignment built into the test development process. And, importantly, schools, teachers and students would have a clearer target for teaching and learning.

Alignment is the agreement between standards and assessments. Problems in alignment can be the result of poorly constructed standards as described above, misdirected assessment items, or both. This study points to the lack of agreement among reviewers as a problem, but more investigation is needed to distinguish if reviewer disagreement is due to lack of training, lack of alignment, or for some other reasons. A simple model can help think about problems raised in this study:

Estimate of Alignment = Standard and Assessment Agreement
+ Standard Ambiguity
+ Assessment Misdirection
+ Reviewer Lack of Agreement
+ Other Inconsistencies

Both analyses discussed in this study produced an estimate of alignment between a set of standards and an assessment. The results estimate the agreement between the standards and an assessment to degree that other factors are not present including standard ambiguity, assessment misdirection, reviewer disagreement, and other sources of inconsistencies. Standard ambiguity includes overlapping standards, overlapping objectives, inadequate coverage of a standard (domain of content) by underlying objectives, lack of clarity in standards and objective statements, statement of processes rather than statement of outcomes (e.g. students will begin computing with fractions), and other issues. Assessment misdirection includes an insufficient number of items to make a judgment on students' proficiency of a standard, too low of complexity compared to standards, insufficient coverage of the content included under a standard, inappropriate emphasis, and other issues. Reviewer lack of agreement includes

insufficient training on the process, insufficient depth of understanding of the standards, lack of content knowledge, inappropriate use of secondary objectives, and fatigue among other issues. Other inconsistencies include, for example, coding errors (mistake in writing down the appropriate objective number) and mistakes on computing results.

Clearly, reviewer lack of agreement is one source of inconsistency that is important to consider in analyzing alignment. Including a larger number of reviewers, averaging results among reviewers, and improving training all are means for reducing the issues related to reviewer lack of agreement. But there are other sources that contribute to inconsistencies among the reviewers that are directly related to alignment issues. These sources include poorly written standards and assessments that do not adequately measure the full intent of the standards. This study identifies the issue of reviewer disagreement and points to the need for alignment analyses to consider more carefully the sources for this disagreement.

In conclusion, this study showed a wide range of reviewer agreement during the process of aligning standards and assessments, with substantial reviewer disagreement about important elements such as correspondence between objectives and items on the assessments. Taking this reviewer disagreement into account changed conclusions about alignment, not only showing weaker alignment than previously demonstrated, but also changing the profiles of alignment about, for example, relative coverage of specific standards. Given the importance of alignment of standards and assessments, as well as the requirement of alignment under current regulations of No Child Left Behind, these results raise red flags about currently used procedures for analyzing alignment. They point to the need for greater clarity in objectives and standards, more extensive reviewer training during the alignment process, and possibly also inspection of items to uncover characteristics that may lead to uncertainty among reviewers.

References

Ananda, S. (2003). Achieving alignment. *Leadership, 33*, 18-21.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues & Practice*, 22, 21-29.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M., (2000). *Alignment of standardized achievement tests to state content standards: a comparison of publishers' and teachers' perspectives*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

Herman, J. L. (2004). The effects of testing in instruction, In Fuhrman, S & Elmore, R. *Redesigning accountability systems for education.* New York: Teachers College Press.

Herman, J. L., Webb, N. M., Zuniga, S. A. (2003). *Alignment and college admissions: the match of expectations, assessments, and educator perspectives.* (CSE Technical Report #593). Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., Webb, N. M., Zuniga, S. A. (2005, April). Measurement Issues in the Alignment of Standards and Assessments: A Case Study. Paper presented at the annual conference of the American Educational Research Association, Montreal.

Koretz, D. M., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Koretz, D. M., Mitchell, K. J., Barron, S. & Keith (1996). *Perceived effects of the Maryland State Assessment Program.* (CSE Technical Report #406). Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lane, S.; Stone, C., Parke, C., Hansen, M. & Cerillo, T. (2000). *Consequential evidence for MSPAP from the teacher, principal and student perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

McDonnell, L. M. & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments.* (CSE Technical Report #442). Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Olson, L. (2003, Spring). Standards and tests: Keeping them aligned. *Research points: Essential information for education policy, 1*(1).

Porter, A. C. (2002). Measuring the Content of Instruction: Uses in Research and Practice. *Educational Researcher, 31*, 3-14.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators*. CPRE Research Report Series. Consortium for Policy Research in Education, Philadelphia, PA.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing.* (CSE Technical Report #566). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stecher, B.; Barron, S.; Chun, T., & Ross, K. (2000). *The effects of the Washington State Education Reform on schools and classrooms* (CSE Tech. Rep. No. 525). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Council of Chief State School Officers and National Institute for Science Education, Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002). *An Analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the American Educational Research Association Annual Meeting. New Orleans, LA.

Webb, N. L. (2005). *Alignment Analysis of Mathematics Standards and Assessments.*

Wilson, M.R & Berenthal, M.W. (2005). *Systems for State Science Assessment.* Washington DC: National Academies Press