Alignment and College Admissions: The Match of Expectations, Assessments, and Educator Perspectives

CSE Technical Report 593

Joan L. Herman, Noreen Webb, & Stephen Zuniga
CRESST / University of California, Los Angeles

April 2003

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

Project 2.3 Indicators of Classroom Practice and Alignment
Joan L. Herman, Project Director, CRESST / University of California, Los Angeles

The work reported herein was supported under the Validity of the Golden State Exam for Use in UC Admissions Project, PR / Award pending, as administered by the University of California Office of the President and the Educational Research and Development Centers Program, Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the University of California Office of the President nor the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

ALIGNMENT AND COLLEGE ADMISSIONS: THE MATCH OF EXPECTATIONS, ASSESSMENTS, AND EDUCATOR PERSPECTIVES

Joan L. Herman, Noreen Webb, & Stephen Zuniga
CRESST/University of California, Los Angeles

Abstract

This study examined the alignment between the Golden State Exam (GSE) in High School Mathematics and the University of California Statement on Competencies in math, exploring the technical quality of the alignment process. UC faculty and high school math teachers (N = 20) from Northern and Southern California rated the math items of the GSE relative to the expectations identified in the UC competency statement, identifying item features related to content and complexity. Raters assigned values for a primary topic, secondary topic, item / topic centrality, depth-of-knowledge, and source of challenge for each item. Agreement within these criteria was the basis of the assessment of alignment. Results showed that there was moderate to strong agreement between faculty and teachers in topic and category identification. Also, there was a moderate relationship between depth of knowledge ratings and item complexity and difficulty based on ratings and student performance. These results suggested that there was an overall good alignment between the GSE and its intended targets and raised methodological issues pertaining to the alignment of standards and assessments.

Introduction

Alignment is a core validity issue in today's standards-based assessment systems. It seems almost self-evident that if standards specify what students should be taught and what and how well they should be learning, and if tests measure what students know and can do, the two need to be closely synchronized. Further, there needs to be a close match between the two if assessment systems are to serve their intended purposes in support of reform. These purposes include: communicating the standards to educators, parents and students; focusing instruction on the standards; providing accurate information to schools and their publics about how students are doing relative to standards; and grounding accountability systems that stimulate improvement and progress toward all students achieving agreed upon standards. Indeed, assessments that are mismatched with standards may be counter-

productive to desired consequences. Because research clearly shows that, faced with accountability, teachers tend to focus their instruction on what's tested (McDonnell & Choissier, 1997; Lane, Stone, Parke, Hansen, & Cerillo, 2000; Stecher, Barron, Chun, & Ross, 2000; Koretz, Mitchell, Barron, & Keith, 1996), a mismatch may encourage attention to low-priority knowledge and skills at the expense of more valued ones. Additionally, it is clear that absent strong alignment, tests cannot yield accurate inferences about students' attainment of, and progress toward, standards, nor about the success of the reform effort.

Despite the importance of the concept, the present state of alignment is weak (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Rothman, Slattery, Vranek, & Resnick, 2000), and sound methodologies for examining and documenting it are just recently emerging. The reasons for this state of affairs are many: the vagueness of many standards; a developmental time frame in which tests precede the development of standards; unreasonable test production schedules that preclude careful study; and the difficulties of reliably assessing important aspects of standards, to name just a few. If we are going to move forward with standards-based accountability and assessment in the service of improved learning, we need to do better. To do so, we also need better methodologies for judging alignment, methodologies that recognize the meaning and complexity of the concept of alignment and that can support broader reform goals.

This paper explores relevant methodologies in a study of the alignment between California's Golden State Examination in High School Mathematics—an honors test being considered as an option in college admissions—and competencies expected of entering freshman in mathematics. The context is the University of California's exploration of alternatives to the SAT-1 for purposes of eligibility and admissions and UC President Atkinson's desire for admissions testing that will have productive impact on high school curriculum and teaching. The case provides an example of an attempt to promote better alignment between the K-12 and higher educational systems and to re-purpose a test. That is, the high school test we investigated was intended to be aligned with the state's K-12 standards, but for admissions purposes, the alignment with University expectations was a key.

The study purposively used multiple raters and separate panels of faculty and high school educators to assess the alignment of the test with University expectations and to explore the agreement between the two groups of raters. Such agreement is very important for both technical and socio-political purposes. From a

psychometric standpoint, sound alignment processes must yield reliable results and, from a socio-political perspective, agreement reflects common understandings that are essential if standards and tests are to serve their intended communication and instructional purposes. The underlying assumption is that the UC *Statement on Competencies* helps high school educators understand what the University expects and that both high school educators and UC faculty share a common interpretation of the meaning of the topics and expectations. With such common understandings, we can be more confident that high schools will know what they are supposed to be teaching students to prepare them for UC. In the absence of common understandings, high school educators may think they are preparing their students for UC expectations, but the specifics of their content teaching may be at variance with the expectations. Moreover, if studying for the test is intended to help students prepare for UC and if test results are supposed to be an indicator of preparation relative to expected competencies, the alignment between the test and the competencies is essential. Thus, the study addresses three general issues:

1. Agreement/Reliability: How can the reliability of the alignment process be assessed? To what extent are high school educators and college mathematics faculty consistent in their judgments of alignment?
2. Alignment: To what extent is the subject test aligned with UC expectations? How does the test developers' classification of items compare with the independent ratings gathered through this study?
3. Implications: What are the implications of the study for future research and practice?

Methodology

In this section we describe the panels that were convened to judge alignment and the tools they used to make the judgment: the high school mathematics test, the UC Statement on Competencies Expected for Entering Freshman and the alignment instrument used for making comparisons between the two. We then summarize training and rating procedures and present an overview of our analysis strategies.

The Raters

As noted above, we convened panels of University mathematics faculty and high school mathematics educators to conduct the study. We solicited recommendations for panelists from UC faculty, department chairs at each campus,

and the California Subject Matter Projects. We looked for individuals who were subject matter experts and experienced in reform and assessment issues in the K-12 educational system. A total of 10 faculty members and 10 high school educators were recruited, and separate panels of each were convened in both Northern and Southern California.

The 10 teachers who rated the exam have a substantial background in K-12 math assessments. They averaged 13 years of teaching and have experience in grading statewide exams. The teachers have worked in developing district standards, written exit exams, assisted the UC system in the development of high school math programs, and worked in the development of the GSE as well as grading the exam. The UC faculty is equally experienced, including professors with more than 30 years of experience in the UCs. Their background in K-12 math assessments include participation in the Mathematics Framework Committee, Stanford-9 math content review board, grading of statewide math exams, the California Math Project, and original members of the committee that formed the Statement of Competencies in Mathematics.

High School Mathematics Golden State Examination

First administered in spring 1998, the Golden State Exam in High School Mathematics (GSE) is intended for students who have completed 2 years of high school algebra and geometry. The exam is administered in two 45-minute sessions during the winter of each academic year. For the 2001 administration, each session consisted of 20 multiple-choice questions and one written-response item, with calculators allowed only during session two. The content is based on the *Mathematics Standards for California Public Schools, Kindergarten through Grade 12* and covers algebra I, geometry, algebra II, and probability and statistics. According to the California State Department specifications for the 2001 examination, test content across these topics is described and distributed as shown in Table 1. Intended processes are summarized in Table 2.

The test questions are developed by teams of teachers, subject matter specialists, and university professors; reviewed for content and bias; field tested; and subjected to additional review before a test becomes operational. With the exception of a few common items that are retained from year to year for purposes of linking and equating results, new test forms are created each year. The multiple-choice questions on the test are machine scored and panels of California math

teachers are convened to score the written response items, using scoring guides that are specially developed for each item. Open-ended items are scored on a four-point scale. Scores on the multiple-choice and written response items then are combined to classify students into one of six performance levels. As noted above, students designated at one of the three highest levels (recognition, performance level 4; honors, performance level 5; high honors, performance level 6) receive recognition as Golden State Scholars.

Table 1

Intended Content*
Winter 2000 High School Mathematics Golden State Exam

1. **Algebra I Standards:** Algebra I develops an understanding of the symbolic language of mathematics using symbolic reasoning and making calculations with symbols. Skills and concepts are used in a wide variety of problem-solving situations (20% of items).
2. **Geometry Standards:** Geometry develops an understanding and use of the mathematics of points, lines, angles, surfaces and solids, including right triangle trigonometry (33% of items).
3. **Algebra II Standards:** Algebra II goes beyond the concepts of Algebra I and Geometry and provides experiences with algebraic solutions of problems in various content areas (35% of items).
4. **Probability and Statistics Standards:** Probability and Statistics develops an understanding of the processing of statistical information, including the study of probability, interpreting data, and fundamental statistical problem solving (12 % of items).

*The test on definition of content is taken from California State Department of Education, *High School Mathematics Blueprint* (2001a). Data on the percentage of items addressing each content area is drawn from *Overview Specifications for High School Mathematics* (CSE, 2001b).

Table 2

Intended Processes*
Winter 2000 High School Mathematics Golden State Exam

1. **Know and understand**, requiring the use of mathematical definitions and / or relationships (22.5% of items).
2. **Interpret**, requiring students to process information by using multiple mathematical relationships (30% of items).
3. **Apply/Analyze**, requiring students to use and analyze mathematical concepts and relationships and apply them (35% of items).
4. **Synthesize**, requiring students to use original thinking to address non-routine situations (12.5% of items).

* The test on definition of each process is taken from California State Department of Education, *High School Mathematics Blueprint* (2001a). Data on the percentage of items addressing each process is drawn from *Overview Specifications for High School Mathematics* (CSE, 2001b).

Table 3 presents the mean, standard deviation, and the KR-20 statistic for the multiple-choice section of the exam. The Mislevy coefficient, indicating the reliability of the combined multiple-choice and open-ended items was .82.

Statement on Competencies in Mathematics Expected of Entering College Students

The *Statement on Competencies in Mathematics Expected of Entering College Students* was developed by a joint task force of representatives from the University of California, California State University, and the California Community Colleges. The UC Academic Senate subsequently adopted it as the institution's formal position on mathematics expectations. The document is intended to provide a clear picture of what mathematics students need to know and be able to do to be successful in college. The first section describes the characteristics and dispositions of students who are well prepared and the range of instructional processes in which students should be engaged to be successful. A second section summarizes the experience and background in technology to which students should be exposed. Section III, the core section for the current study, describes areas of mathematical content that are: 1) essential for all entering college students; 2) desirable for all entering college students; 3) essential for college students to be adequately prepared for quantitative majors; and 4) desirable for college students who intend to declare quantitative majors (see Appendix: Competency Topic List). (The full statement can be found at www.ucop.edu/senate/index2.html)

Alignment Rating Instrument

The Alignment Rating Instrument was adapted from one developed by Norman Webb (1997). It asked reviewers to examine each item on the examination and:

Table 3

Descriptive Statistics and Multiple-Choice Reliability Analysis

Number of MC items:	40
Multiple Choice mean:	21.73
Std. Deviation:	6.82
N =	49,596
KR-20:	0.82

1. Identify the content topic(s), if any, from the *Statement on Competencies* to which each item corresponds. Raters could identify both a primary and secondary topic, as appropriate.
2. Rate the centrality of the item to the topic it addresses, using the following rating scale
   (1) Within the topic area, but not essential for students
   (2) Within the topic area, of moderate importance
   (3) Within the topic area, of central importance
3. Judge the depth-of-knowledge level of each assessment item, using the following levels:
   (1) Recall and Reproduction
   (2) Skills and Concepts
   (3) Problem Solving and Strategic Thinking
   (4) Extended Thinking
4. Judge whether there is a source-of-challenge issue with an item, which was only used to identify items where item difficulty is inadvertently introduced by characteristics unrelated to the targeted mathematical knowledge. That is, the item is difficult not because of the mathematics it involves but because of other unrelated factors such as excessive reading demands, confusing language, cultural bias, etc.

Procedures

As noted, separate panels of high school educators and University faculty were convened in Northern and Southern California. Before the meetings, participants were informed of the general goals of the activity and were sent a *Statement on Competencies* to review. At the meetings, participants were given additional orientation to the project and its goals and then introduced to the rating instrument and process.

Participants reviewed the specific written guidance provided on each of the rating criteria. There was extended discussion about what constituted the various levels of depth of knowledge and the hierarchy these levels represent. As described by Webb (op. cit.), the hierarchy is based on two factors, the mathematical

sophistication of the item and the likelihood that students were familiar with the problem type through prior instruction. The mathematical sophistication of the item depends on the abstractness of the problem, the amount of mathematics that needs to be used, the number of mathematical principles to be employed, the problem novelty, and the need to extend or produce original findings. However, some assessment items may look challenging to a novice but in fact represent a low depth-of-knowledge level because the knowledge required to solve the item is commonly known, and students are likely to have had the opportunity during normal instruction to routinely (habitually) solve such items. Anything that was considered routine or algorithmic in this sense was considered level one depth of knowledge.

After practicing using the coding scheme, sharing answers, and reaching reasonable levels of agreement, panel participants then embarked on individual ratings of each item on the test. After all ratings had been completed, a debriefing session was held in which participants were invited to respond to the following questions:

1. General reactions?
2. What did you think of the balance of representation of content areas?
3. What did you think of the balance of representation in the depth of knowledge reflected on the test? What depth of knowledge represents college level work?
4. Did the competency statement provide you enough information to make the judgment?

Analysis Plan and Decisions

The analysis plan considered appropriate ways to assess rater consistency as well as decisions about when to consider an item aligned with a particular topic or category. In addition to reporting descriptive information showing exact agreement among raters, we calculated kappa coefficients for categorical ratings (mathematics topic and category assignment) and dependability coefficients for ratings that had inherent quantitative meaning (item complexity, as indicated by raters giving both a primary and a secondary topic rating or only a primary rating; depth of knowledge, centrality of the item for measuring a particular topic).

Agreement analyses. To examine the agreement among raters for the categorical ratings, we calculated kappa coefficients of agreement (Cohen, 1960; Fleiss, 1971). Because the kappa coefficient takes into account chance agreement

among observers, it is preferred over other summary indices such as exact percent of agreement (see Watkins & Pacheco, 2001).

To examine the agreement among raters for the quantitative ratings, we conducted generalizability analyses with items crossed with raters. The items were considered the object of measurement, and variation among raters was considered error variation. Each generalizability analysis produced an estimated index of dependability, a reliability-like coefficient that showed the consistency of raters in coding attributes of the items. The index of dependability provides information on the absolute level of rater agreement in their ratings of items, not only the consistency of raters in their assessment of the relative standing of items, on a particular item attribute such as depth of knowledge (see Brennan, 2001; Shavelson & Webb, 1991).

Yardsticks for determining when a topic or category was covered. To examine the alignment between the test items and the *Statement on Competencies*, we had to have decision rules to determine when an item was to be considered a match to a particular topic or category in the statement. That is, how many raters should agree before the match should be credited? Since this is an exploratory study, we decided to consider both a lax and a more stringent standard—a bare majority who agreed that an item measured a particular topic or category (more than half, or 55% and above) and a clear majority (70% and above).

Decision to consider either primary or secondary topic identified. Since our methodology enabled raters to identify both a primary and a secondary topic for an item, a second consideration in determining alignment was the issue of which of these ratings needed to agree for a determination that raters were in agreement. We used relaxed criteria, because of the nature of both our training processes and the *Statement on Expectations*. That is, while our training process included general rules for defining primary and secondary topics, because of time pressures, we did not train extensively on assuring the differentiation between the two. Further, and more importantly, the *Statement on Expectations* was written to provide guidance to high school educators and students about what students should know and be able to do, and not to support the purpose for which we were using it. The topic categories were written to be descriptive and inclusive rather than mutually exclusive. Topics and subtopics are listed, but not operationally defined. Therefore, we decided to combine the primary and secondary ratings given by each rater to each item and use whichever one, if either, agreed more strongly with the ratings given by other raters.

Results

We begin the results section with description data relevant to study variables. Results are then presented about the rater agreement we found and substantive findings with regard to the alignment of the GSE with the *Statement on Competencies.*

Descriptive Information

Descriptive data about the study's central variables are summarized relative to the California Department of Education's test blueprint (2001a). That is, we used the state's determination of content and category for each item to describe the numbers of topics and categories assigned to each item, and the complexity, depth of knowledge and centrality of each.

Mathematics topic and category ratings. Table 4 gives the number of mathematics topics and the number of mathematics categories that raters gave, on average, for the items within each of the content areas the test was designed to assess, using the content classification given by the California State Department of Education (CDE; 2001a). The number of topics and the number of categories include

Table 4

Number of Mathematics Topics and Categories Selected Per Item by Content Area

	Number of Topics Selected Per Item[b]			Number of Categories Selected Per Item[b]		
Content Area[a]	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Algebra I (9 items)	5.9	3.2	3-11	2.4	1.3	1-5
Geometry (13 items)	6.9	1.8	3-9	2.8	1.2	1-5
Algebra II (16 items)	5.9	3.2	3-13	2.8	1.0	1-4
Probability / Statistics (6 items)	7.5	3.2	3-12	3.5	2.0	2-7

[a]Classifications determined by the California State Department of Education (2001a).

[b]Includes both primary and secondary topic / category ratings. Raters could choose from among 58 topics, classified in 10 categories.

both primary and secondary ratings. As can be seen in Table 4, the 20 raters assigned a fairly large number of topics to each item (e.g., nearly 6 topics, on average, per item classified by the California Department of Education as Algebra I). The number of categories assigned to each item was smaller, but was substantial compared to the total number of categories raters could choose from. The number of topics and categories chosen by the high school educators tended to be slightly higher than the number of topics and categories chosen by the University faculty. However, because the differences between rater groups were not statistically significant, the information is not presented separately here by rater group.

Table 5 shows the topics chosen most frequently for the items in each of the four content areas designated by the CDE (2001b). Table 6 presents the categories chosen most frequently for each item. While not a perfect match, there seems to be general agreement between the topics and categories chosen most frequently by study raters and CDE's classification of items to content areas.

Item complexity: Primary and secondary topic ratings. Raters had the option of assigning both a primary topic rating and a secondary topic rating to an item. Items with a large number of secondary ratings were considered more complex than were items with a small number of secondary ratings. Table 7 gives the average number of raters (maximum of 20) who assigned secondary topic ratings per item. As can be seen in Table 7, the items classified by the CDE (2001b) as Geometry were the most complex: slightly more than half of the raters (12.4 out of 20), on average, perceived the items as sufficiently complex to merit assigning a secondary topic rating in addition to the primary topic rating. Using that same yardstick, the items representing Probability/Statistics were the least complex: fewer than half of the raters (7.3 out of 20), on average, perceived the items as sufficiently complex to merit assigning a secondary topic rating.

Depth of knowledge and centrality. Raters were asked to judge the depth of knowledge level (on a scale of 1 to 4, with 4 representing highest depth of knowledge) and the centrality of the item to the topic it addressed (on a scale of 1 to 3, with 3 representing the highest centrality). Table 8 gives the average depth of knowledge and centrality across the items in the four content areas given by the CDE (2001b). Most raters perceived most items as requiring a moderate level of depth of knowledge: between recall and reproduction (level 1) and skills and

Table 5

Mathematics Topics Selected by the Greatest Number of Raters (Per Item) by Content Area

Content Area[a]	Topic	Number of Items
Algebra I	Solutions of linear equations and inequalities	2
	Linear functions	2
	Curriculum prior to Algebra 1	2
	Translation from words to symbols	1
	Solutions to systems of equations, and their geometrical interpretation	1
	Solutions of quadratic equations	1
Geometry	Distances, areas, and volumes, and their relationship with dimension	4
	Similarity	4
	Angle measurement	3
	Pythagorean Theorem	2
Algebra II	Exponential functions	3
	Counting (permutations and combinations, multiplication principle)	3
	Quadratic and power functions	2
	Solutions of linear equations and inequalities	1
	Powers and roots	1
	Solutions to systems of equations, and their geometrical interpretation	1
	Polynomials	1
	Exponential functions	1
	Families of Functions and Their Graphs	1
	Solutions of quadratic equations	1
	Function notation	1
Probability / Statistics	Counting (permutations and combinations, multiplication principle)	2
	Using lines to fit data and make predictions	1
	Expected value	1
	Presentation and analysis of data	1
	Curriculum prior to algebra 1	1

[a]Classifications determined by the California State Department of Education (2001a).

Table 6

Mathematics Category Selected by the Greatest Number of Raters (Per Item) by Content Area

Content Area[a]	Category	Number of Items
Algebra I	Variables, equations, and algebraic expressions	5
	Families of functions and their graphs	2
	Probability	1
	Curriculum prior to Algebra 1	1
Geometry	Geometric Concepts	10
	Variables, Equations, and Algebraic Expressions	3
Algebra II	Families of functions and their graphs	8
	Probability	4
	Variables, Equations, and Algebraic Expressions	4
Probability / Statistics	Probability	5
	Data Analysis and Statistics	1

[a]Classifications determined by the California State Department of Education (2001a).

Table 7

Number of Raters Who Assigned a Secondary Mathematics Topic Rating Per Item

Content Area[a]	Number of Raters Who Assigned a Secondary Topic Rating		
	M[b]	*SD*	*Range*
Algebra I (9 items)	8.4	4.7	2-13
Geometry (13 items)	12.4	3.8	6-18
Algebra II (16 items)	9.9	3.9	1-14
Probability / Statistics (6 items)	7.3	3.1	2-11

[a]Classifications determined by the California State Department of Education (2001a).
[b]Maximum possible = 20 raters.

Table 8

Average Depth of Knowledge and Centrality of Items for All Raters

	Depth of Knowledge		Centrality	
Content Area[a]	*M*	*SD*	*M*	*SD*
Algebra I (9 items)	1.40	.25	2.76	.30
Geometry (13 items)	1.67	.21	2.59	.34
Algebra II (16 items)	1.68	.20	2.70	.41
Probability / Statistics (6 items)	1.68	.28	2.64	.34

[a]Classifications determined by the California State Department of Education (2001a).
[b]Includes both primary and secondary topic / category ratings.

concepts (level 2). On average, raters perceived Algebra I items as requiring less depth of knowledge than items classified as Geometry, Algebra II, or Probability / Statistics. The ratings of centrality of the item to the topic it addressed were high (of central importance) for nearly all items.

Source of challenge. Very few "source of challenge" issues were identified. Therefore, this variable was dropped from further analysis.

Rater Agreement

This section examines both agreement among raters when assigning mathematics topic, mathematics category, depth of knowledge, and item centrality ratings to each item, and the extent to which the two categories of raters (university faculty, secondary school teachers) agreed with each other.

Classification of test items by topic and topic category. To examine the extent to which raters agreed on the topic that an item measured, we first looked at the distributions of agreement across raters on each item. Table 9 shows the distribution of agreement on the topic ratings made by faculty, by high school educators, and by the full set of faculty and high school educators. For faculty, on 36 out of 42 items, a

Table 9

Distribution of Rater Agreement on the Topic Assignments

Proportion of Raters Agreeing on Topic Assignment[a]	Rater Category		
	Faculty	High School Educators	Faculty and High School Educators Combined
1.0	4[b]	13	7
.9	12	4	12
.8	8	6	5
.7	11	4	6
.6	5	7	5
.5	2	3	4
.4	4	3	1
.3	1	2	2
.2	0	0	0
.1	0	0	0
.0	0	0	0

[a] Proportions are rounded to the nearest tenth. At least 55%, or 11 or 20 raters, would be rounded up to 60% in the table.
[b] Number of items.

majority (at least 60%) agreed on the topic classification. On only four of 42 items was the level of agreement less than 50%. There was total agreement on four items and high agreement (80% or higher) on half of the items. Similarly, for high school educators, on 33 out of 42 items, a majority of raters agreed on the topic classification. On only six items was the level of agreement less than 50%. Teachers showed perfect agreement on 12 items, and high agreement (80% or above) on 24 items. On 35 out of 42 items the majority of the combined set of faculty and high school educators agreed on how the items should be classified; on only four items was there less than 50% agreement. On 28 items, at least 70% of the group agreed on the classification. As noted above, these two different classification standards of agreement—more than 50% or at least 70%—were used in the analyses of item content and content coverage of the test, as will be described further below.

It should be noted that the seven items for which less than a majority of raters agreed on the topic classification were distributed across all of the content areas. Furthermore, the level of agreement for some items may have been underestimated

in two ways. First, raters sometimes selected different topics that were very similar in wording. For example, one item was rated as measuring "solutions to systems of equations, and their geometrical interpretation" by one set of raters and as measuring "solutions to quadratic equations, both algebraic and graphical" by another set of raters. Had these topics been declared to be the same, this item would have yielded 80% agreement among raters instead of 50%. Since both were separate topics listed in the *Competency Statement*, we chose to retain them as separate in our analysis. However, future studies consider additional validity checks to see whether raters are treating topics as different, and/or additional advance check to consolidate topics that may not be well differentiated Second, some raters sometimes selected a topic within a category as the best match for an item while other raters selected the category itself. For example, one item was rated as measuring "interpretation of graphs" by one set of raters and as measuring the overall category for that topic "families of functions and their graphs" by another set of raters. Had the category and the specific topic within the category been declared to be the same for purposes of measuring agreement among raters, the level of agreement for that item would have been 80% instead of 45%. Making distinctions between topics with very similar labels and between topic and category designations affected the measurement of rater agreement for all items, not only those with the lowest levels of agreement.

Table 10 gives the distributions of agreement of faculty, teachers, and faculty and teachers combined for the mathematical categories into which each item should be assigned. When assigning items to mathematical categories, raters agreed highly. On every item, a majority of faculty raters agreed on the category the item measured. On all items except three, a majority of teacher raters agreed on the category the item measured. Combining faculty and teacher raters, the majority of raters agreed on their assignments of items to categories, with the exception of one item. The results for interrater agreement, then, showed substantial agreement among raters about which specific topic an item measured and which mathematical category applied to an item.

To provide summary statistics about interrater agreement that correct for chance, we calculated kappa coefficients of interrater agreement for raters' topic assignments (57 specific topics) and category assignments (10 categories). Table 11 presents kappa coefficients for all 42 items, for the 40 multiple-choice items, for the two written items, and for the items designed by the California State Department of

Table 10

Distribution of Rater Agreement on the Category Assignments

Proportion of Raters Agreeing on Category Assignment	Rater Category		
	Faculty	High School Educators	Faculty and High School Educators Combined
1.0	16[a]	22	20
.9	7	6	8
.8	10	4	5
.7	5	3	7
.6	4	4	0
.5	0	1	0
.4	0	2	2
.3	0	0	0
.2	0	0	0
.1[b]	0	0	0

[a]Number of items.
[b]With 14 categories to choose from, the lowest level of agreement among raters that was statistically possible was .07.

Table 11

Assignment of Items to Topic or Category: Kappa Coefficients of Inter-rater Agreement

Item Set[c]	Number of Items	Topic Ratings[a]		Category Ratings[b]	
		Faculty	Teachers	Faculty	Teachers
All items	42	.55	.58	.71	.74
Multiple choice items	40	.55	.58	.69	.73
Written items	2	.40	.46	.99	.99
All Algebra items[d]	24	.55	.62	.66	.73
Algebra I	9	.58	.60	.66	.73
Algebra II	16	.50	.60	.66	.65
Geometry	13	.52	.49	.49	.47
Probability / statistics	6	.35	.43	.36	.60

[a]Raters could choose among 58 specific mathematics topics.
[b]Raters could choose among 10 categories of mathematics topics.
[c]Classifications determined by California Department of Education (2001a).
[d]One written item was classified by the California Department of Education as measuring both Algebra I and Algebra II.

Education (CDE, 2001b) to correspond to Algebra I, Algebra II, Geometry, and Probability / Statistics. According to the guidelines suggested by Watkins & Pacheco (2000; see also Cicchetti, 1994; Fleiss, 1981), kappa coefficients greater than .75 indicate excellent agreement; values between .60 and .75 indicate good agreement; values between .40 and .60 indicate fair agreement; and values below .40 indicate poor agreement. Table 11 shows moderate agreement among faculty and among high school teachers about which specific mathematical topic an item measured. Agreement about which mathematical category to assign an item was higher, generally in the range of good to excellent. The levels of agreement were similar for faculty and teachers.

Not only were levels of agreement among faculty raters and among teacher raters similar for topic ratings and for category ratings, the particular topics assigned by the two groups of raters were very similar to each other. For 36 out of 42 items, the topic assigned to an item by the largest number of faculty raters was the same topic assigned to an item by the largest number of teacher raters. For the remaining six items, the topic assigned to an item by the largest number of raters in one group (e.g., faculty) was assigned by a substantial number of raters in the other group (e.g., teachers). So, even among these six items, there was considerable overlap in the topic assignment between faculty and teachers. It should be noted that these six items were not concentrated within any particular mathematics content area but were spread across the Algebra II, Geometry, and Probability / Statistics dimensions generated by the California State Department of Education.

Faculty and teachers agreed even more strongly on the category ratings for each item. For 40 out of 42 items, the category assigned to an item by the largest number of faculty raters was the same category assigned to an item by the largest number of teacher raters. These results show that there was general agreement between the two groups of raters (faculty, teachers) on the classification of content on the test.

Item complexity: Primary and secondary ratings. Each rater was asked to assign a primary topic to each item and, if appropriate, also a secondary topic. As described earlier, items assigned both a primary topic and a secondary topic are considered more complex than items assigned only a primary topic. To examine rater agreement on item complexity, we first examined the degree of agreement among raters on whether they assigned only a primary topic, or both a primary and secondary topic to an item. Table 12 gives the distribution of agreement among

Table 12

Distribution of Rater Agreement on Item Complexity: Assignment of a Secondary Topic Rating

Proportion of Raters Agreeing on Whether an Item Required a Secondary Topic Assignment	Rater Category		
	Faculty	High School Educators	Faculty and High School Educators Combined
1.0	5[a]	1	2
.9	3	6	6
.8	8	8	2
.7	8	12	9
.6	8	8	21
.5[b]	10	7	2

[a] Number of items.
[b]With two options to choose from (assigning both a primary rating and a secondary rating, or only a primary rating), the lowest level of agreement that was statistically possible was .50.

raters on whether they assigned both primary and secondary topics to an item or only primary items. The level of agreement was fairly low. On only a small number of items did raters agree 100% on whether the item required a secondary rating (5 items for faculty raters, 1 item for high school teacher raters, and 2 items for the pooled set of raters). Most items generated considerable disagreement about whether a secondary topic rating was required. Second, we assessed agreement among raters on item complexity using generalizability theory. The estimated coefficients of dependability were quite low, ranging from .14 to .18 for faculty raters, teacher raters, and all raters combined.

Depth of knowledge and item centrality. In addition to assigning a topic (primary, secondary) for each item, raters were asked to indicate the depth of knowledge required for each item and the centrality of the item to the topic it addressed. To examine interrater agreement for the depth-of-knowledge and item centrality ratings, we first give the distribution of levels of agreement for faculty, teachers, and for all raters combined for depth of knowledge (Table 13) and for item centrality (Table 14). The levels of agreement among raters on the depth of knowledge required by each item were substantial. On 35 items, the majority of faculty raters agreed on the level of depth of knowledge required by the item; on 36 items, the majority of teacher raters agreed on the level of depth of knowledge

Table 13

Distribution of Rater Agreement on the Depth of Knowledge

Proportion of Raters Agreeing on Depth of Knowledge	Rater Category		
	Faculty	High School Educators	Faculty and High School Educators Combined
1.0	2[a]	1	2
.9	8	3	2
.8	6	9	16
.7	10	10	7
.6	9	13	7
.5	7	5	8
.4	0	2	0
.3[b]	0	0	0

[a]Number of items.
[b]With 3 ratings to choose from, the lowest level of agreement among raters that was statistically possible was .33.

required by the item; and on 34 items, the majority of total raters agreed on the level of depth of knowledge required by the item.

A similar picture emerged for centrality of the item to the topic it addressed. As can be seen in Table 14, on most items, the majority of raters agreed on the centrality of an item. On 37 items, the majority of faculty raters agreed on the level of centrality of an item to the topic it measured; on 36 items, the majority of teacher raters agreed on the level of centrality of an item to the topic it measured; and on 33 items, the majority of total raters agreed on the level of centrality of an item to the topic it measured.

Agreement among raters on depth of knowledge and item centrality was also assessed using generalizability theory. The estimated coefficients of dependability for the depth of knowledge ratings ranged from .30 to .31 for ratings given by faculty raters, teacher raters, and all raters combined. The estimated coefficients of dependability for the item centrality ratings ranged from .05 to .10 for ratings given by faculty raters, teacher raters, and all raters combined. The low coefficients of dependability for the depth-of-knowledge and the item centrality ratings are due in large part to the restriction of range in the ratings across items (small item variance). Restriction of range in the ratings across items depresses estimates of dependability,

Table 14

Distribution of Rater Agreement on Item Centrality

Proportion of Raters Agreeing on Item Centrality	Rater Category		
	Faculty	High School Educators	Faculty and High School Educators Combined
1.0	1[a]	2	1
.9	7	7	10
.8	15	11	9
.7	9	9	11
.6	5	7	6
.5	4	6	5
.4	1	0	0
.3[b]	0	0	0

[a]Number of items.
[b]With 3 ratings to choose from, the lowest level of agreement among raters that was statistically possible was .33.

similar to the effect of restriction of range on correlation coefficients. To examine the range of ratings across items, we calculated the mean depth of knowledge rating and the mean item centrality for each item over the 20 raters (faculty and teachers combined) and then examined the mean and standard deviation of those rating means across the 42 items. The variation of mean depth of knowledge ratings across the 42 items was fairly small (SD = 0.35, M = 1.59), showing that raters tended not to use the whole four-point scale when assigning ratings. The variation of mean item centrality ratings was extremely small and the mean was high (SD = 0.17, M = 2.67), showing that most of the item centrality ratings were at the high end of the three-point scale. Due to the extreme restriction of range for the item centrality ratings—the vast majority of raters assigned high ratings to nearly all items—item centrality was dropped from all further analyses.

While the previous section showed that the two-rater groups (faculty, teachers) agreed highly on the classification of content according to topic ratings, some disagreement between the groups emerged on the depth of knowledge ratings assigned. On the average, across the 42 items, high school educators assigned significantly higher depth-of-knowledge ratings than did UC faculty (M = 1.7 vs. M = 1.5 on a scale of 1 to 3; $t(18) = 2.21$, $p = .05$). In other words, high school educators

tended to see the items as requiring a greater depth of processing than did UC faculty. High school educators gave higher depth of knowledge ratings than did UC faculty on 30 items. However, the differences were statistically significant for only three items.

Alignment Between Test and *Statement on Competencies*

The concept of alignment involves a number of dimensions in the match between a set of standards, such as the *Statement on Competencies,* and a test. Among them, as addressed below, are:

- Relevance of items to standards: To what extent does each of the items on the test match content found in the standards?
- Comprehensiveness of standards coverage: To what extent is each of the topics or categories articulated by the *Statement on Competencies* addressed by the test?
- Balance of standards coverage: to what extent do the test items equally represent the various topics and categories addressed by the test? Is the balance purposeful and appropriate to test purpose?

Relevance of items to standards. All items except one were viewed by all raters as occurring on the *Statement on Competencies Topic List* (See Competency Topic List in Appendix). The sole exception was one item that was considered by the majority of raters to measure a mathematics topic occurring in the curriculum prior to algebra. By this simple criterion, the test items indeed reflect the standards in the sense of being relevant to them.

Comprehensiveness of standards coverage. Figure 1 shows the proportion of topics considered essential for all entering freshman that raters judged as being measured by a specific item on the test. Using the more liberal 55% rater agreement criterion, the test is most comprehensive in addressing the category of Variables, Equations, and Algebraic Expressions, roughly corresponding to Algebra I: 75% of the topics in that category were measured by an item on the test. Half the Geometric Concept topics and 40% of the Families of Functions and Probability topics, respectively, also are addressed on the test. Only 25% of the Data Analysis and Statistics topics receive attention.[1] Finally, Argumentation and Proof received no attention on the test. Patterns are similar for the 70% agreement level, except

[1] Note that percentages do not total 100% because of items on which there was not sufficient agreement to classify the items.

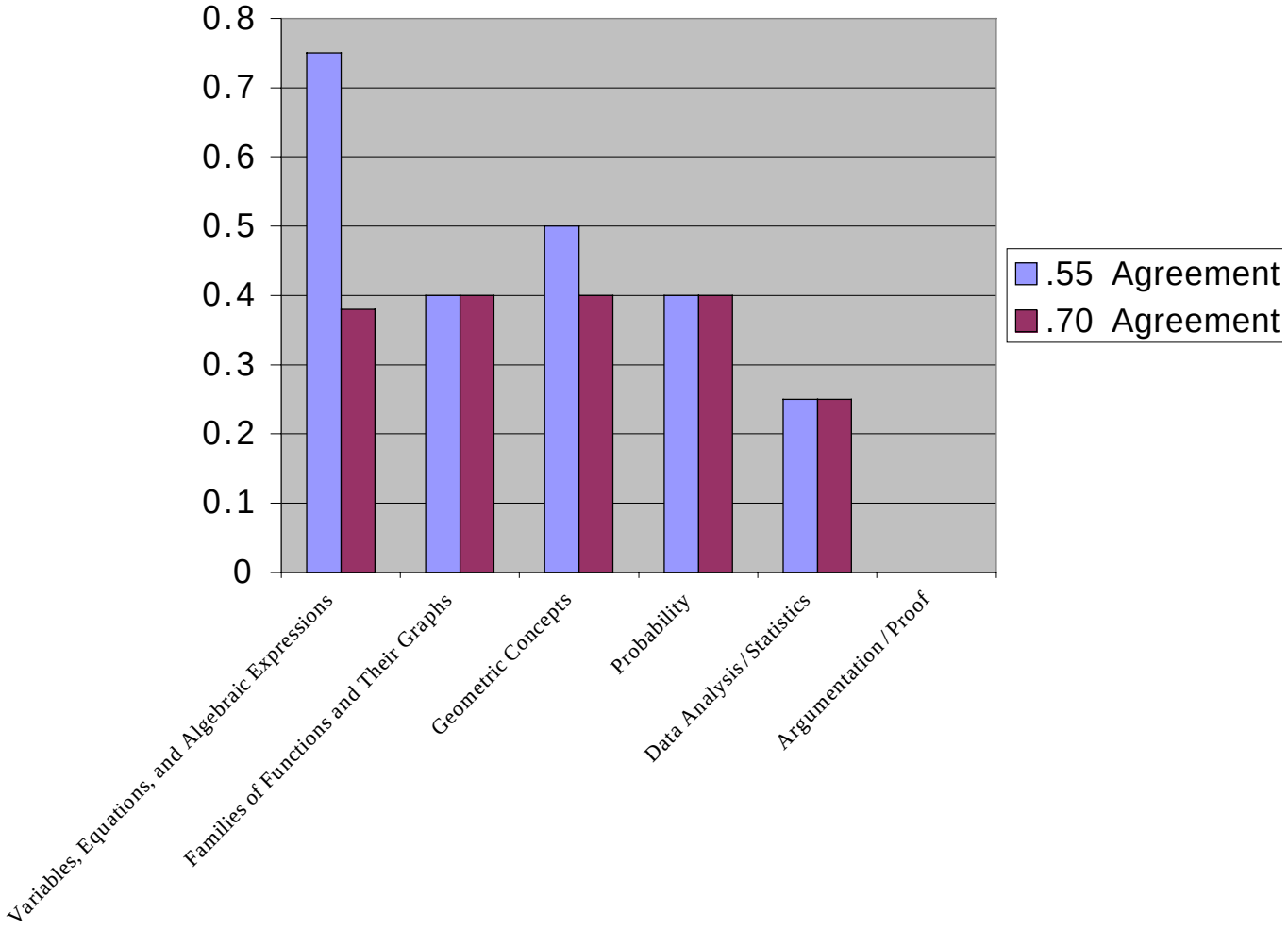

Figure 1. Comprehensiveness of content coverage: Proportion of topics in each category addressed by test. (Note that percentages do not total 100% because of items on which there was not sufficient agreement to classify the items.)

coverage of the category of Variables, Equations and Algebraic Expressions is reduced.

Across all categories of topics considered essential for all entering freshman, only about half of the topics corresponded to an item on the test (46% of the topics at the 55% rater agreement level). The remaining 54% of the topics did not correspond to an item on the test. The topics that corresponded to an item on the test and those that did not were distributed similarly across the categories of variables, equations, and algebraic expressions; families of functions and their graphs; geometric

concepts; probability; and data analysis and statistics. None of the argumentation and proof topics were judged by raters as appearing on the test.

Although the results are not displayed here, the test gives little attention to topics not considered essential for all entering freshman. For example, at the 55% rater agreement level, only 12% of the topics considered essential for quantitative majors correspond to an item on the test. Furthermore, the test gives no attention to topics considered desirable for entering freshman or to topics considered desirable for quantitative majors. Given the original purpose of the test, this pattern of coverage is as might be expected. The test is designed for 11th-grade students (who take the test during the winter), who likely would have only completed Algebra II, and the specifications for the test focuses on Algebra I, II, Geometry, and Probability and Statistics. These findings, however, highlight some of the problems of trying to use a test developed for one purpose for another. That is, the test was not designed to assess advanced mathematics topics the University considers desirable and indeed the test does not address these topics.

Balance of standards coverage. Figure 2 shows how test items are distributed across the various topic categories that the *Competency Statement* deems essential for all entering freshmen. Consistent with the specifications for the test, all of the test items address topics in either Variables, Equations and Algebraic Expressions; Families of Functions and Their Graphs; Geometric Concepts; Probability; or Data Analysis and Statistics. At the 55% rater agreement level, Geometric Concepts receive relatively most attention of the test, with nearly 30% of the items addressing topics in this category. Families of Functions draw 21% of the items, and Variables, Equations and Algebraic Expressions and Probability draw 14% and 10% of the items respectively. Only one item (2% of the test items) addressed data analysis. No item addressed argumentation and proof. Patterns are similar at the .70 agreement level.

It is of interest to note the high agreement between our panelists' ratings of coverage and those that occur in the state's specifications for the test, the High School Mathematics Blueprint. As noted above, the specifications give attention to the major content areas of Algebra I standards, Geometry standards, Algebra II standards and Probability and Statistics standards. Their definition of Geometry standards is equivalent to Geometry Concepts defined by the *Statement on Competencies* and those of Probability and Statistics are roughly equivalent to Competencies in Probability and Data Analysis and Statistics. Similarly, the first two

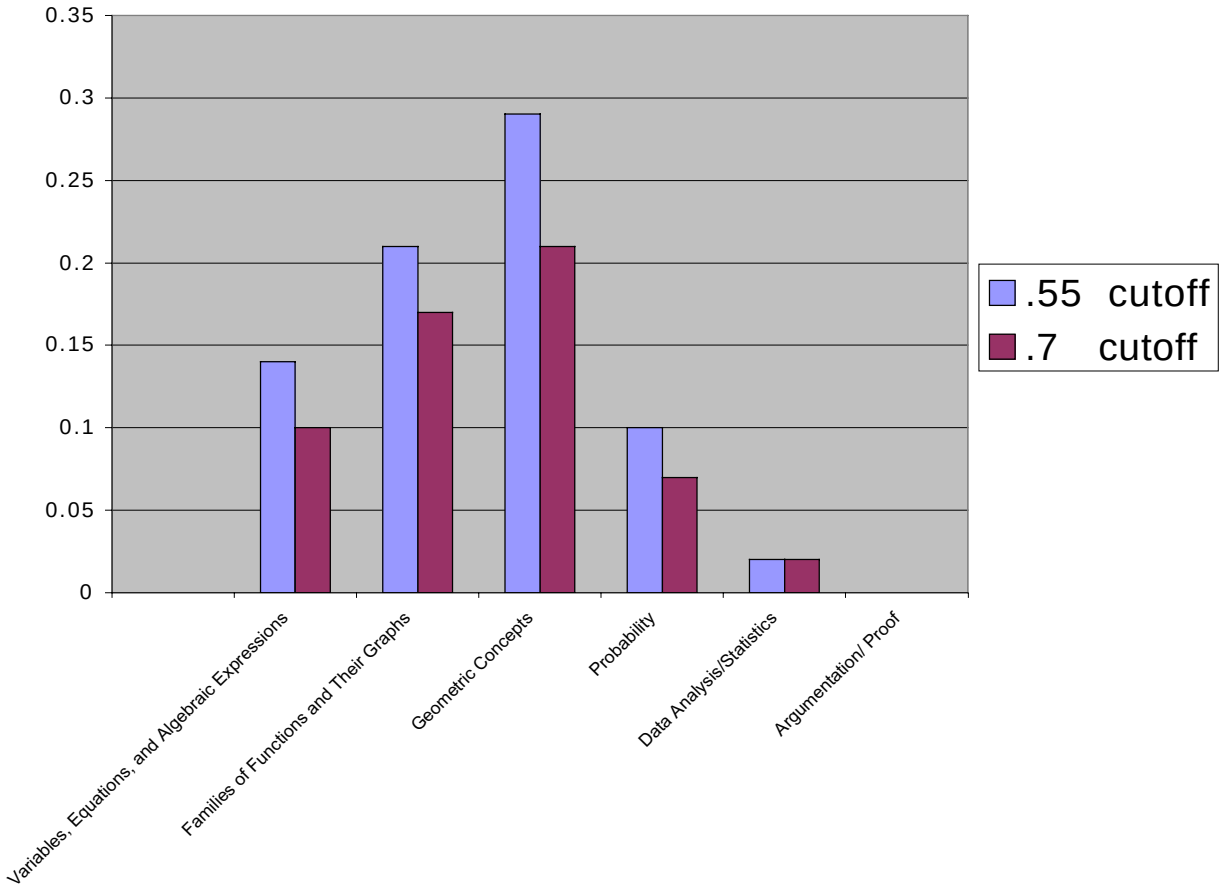

Figure 2. Balance of coverage. Proportion of total test addressing each topic category. Topics considered essential for all entering freshmen. These were items that at least half the panelists had identified as having two topics.

topic categories in our study are similar to Algebra I and Algebra II standards. Figure 3 compares the proportion of items covering each domain based, respectively, on the state's specifications and the findings of this study. Note, however, that the state specifications include all items on the test, while the current study includes only those items on which at least 55% of the raters agreed on the topic or category assignment. From the perspective of the test developers, then, it appears that the current study's raters had most difficulty coming to consensus on items the developers viewed as assessing Algebra I and II. Beyond these differences, the percentages are very similar overall, and spot-checking of individual items, furthermore, shows agreement in topic classification across both sources. From whichever vantage point, Data Analysis and Statistics appear to get relatively little attention.

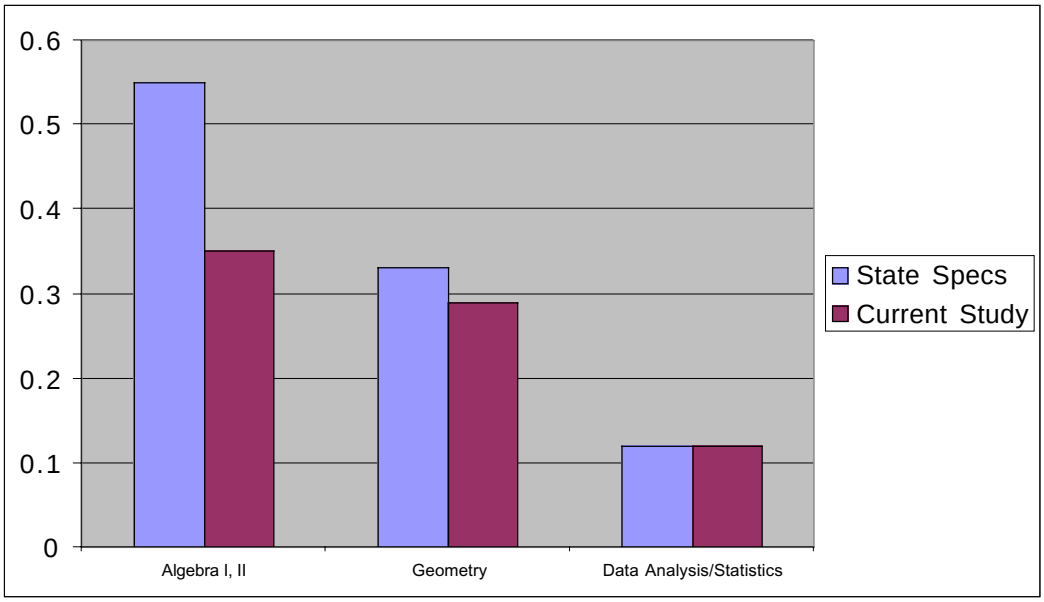

Figure 3. Comparison of Content Classification State Specification Compared to Current Study Results. Note: *Current Study* results do not total 1.00, because of items on which there was insufficient rater agreement to classify content.

Results on depth of knowledge. It is difficult to address issues of the alignment between the University's expectations and the intellectual demands of the test, because the *Statement on Competencies* does not lay out a specific set of expectations in this domain. In our study, the bulk of items on the test seem to address depth of knowledge at level 1 or level 2, and few were rated as level 3. Level 1 denotes the recall of fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. Such problems include one-step, well defined, or straight algorithmic procedures, or problems that ask students to follow a set procedure (like a recipe), or perform a clearly defined series of steps. Level 2 requires engagement of some mental processing beyond recalling or reproducing a response, such as requiring students to make some decisions as to how to approach the problem or activity. Level 3 requires reasoning, planning, using evidence, and a higher level of original thinking; the demands are abstract and complex, and the problems must be novel.

The mean depth of knowledge rating across all items on which there was at least 55% rater agreement was 1.57. Mean depth of knowledge for each category ranged from 1.27 for probability to 1.66 for Geometric Concepts. Differences in mean depth of knowledge across the categories were not statistically significant.

The extent to which items were judged as addressing both a primary topic and a secondary topic provides another window on item complexity. Items that require knowledge of more than one mathematical concept may be more complex than those that require topic knowledge of only one topic, particularly if those topics draw from different domains of mathematics. As noted earlier, a slight majority of the items on the test (57%) were viewed as addressing both a primary and a secondary topic. However, only a third of the items on the test drew on topics that crossed mathematics domains—for example, an item whose response required knowledge of a topic from Variables, Equations and Algebraic Expressions and of a topic from Geometric Concepts.

Panelists' general reactions to issues of alignment. Panelists reacted favorably to the test, although they felt it could be more balanced in terms of its coverage. In particular, a number of panelists were concerned about an over-representation of quadratic equations and inequalities and the inattention to other areas, such as number theory and proofs. Several panelists recommended greater attention to topics that were deemed *desirable* for entering freshman, and a few university faculty in particular expressed some disappointment with the open-ended items. They felt that the problems should be more complex and require mathematical explanation.

The depth of knowledge represented on the test was also a point of contention. There was little consensus on what was the appropriate level of complexity—some thought the test was about right, others thought the test items should be more complex. Faculty members tended to favor greater depth; high school panelists seemed torn between high expectations and their concern that the students they actually taught could function at this high level. There also was discussion about the difficulty of conceptualizing depth-of-knowledge or other constructs reflecting mathematical complexity and processing demands. In this regard, panelists made recommendations for improving the depth of knowledge criterion used in the study and suggested a more sensitive scale that spread the current level 2 ratings across two separate levels, to differentiate between items requiring some and more sophisticated mathematical applications.

The discussion about the balance of content representation also brought to the fore during both faculty panels the issue of the adequacy of the current *Competency Statement* for making such judgments. There was general concern that the current *Competency Statement* is not adequate for judging the alignment of the test, in that the statement does not articulate a clear set of standards and was not developed for this

purpose. Instead, the statement presents a list of topics, coupled with process expectations that outline the types of teaching and learning experiences in which students should be engaged. It does not directly couple content with performance standards.

Relationships Among Item Features and Rater Agreement

This section examines the relationships among rater agreement and various features of the items (e.g., depth of knowledge, item complexity as measured by the number of raters assigning a secondary topic rating to an item, item difficulty). Table 14 presents the correlation among these variables for all raters (faculty and teachers combined) and for faculty and teachers separately.

Correlation between rater agreement and item features. The first row in each matrix in Table 15 concerns the relationship between rater agreement and various features of the items. First, the level of rater agreement for an item was negatively correlated with the average depth of knowledge for that item (for all 20 raters, the correlation was statistically significant; for faculty and teachers analyzed separately, the correlations were negative but did not reach statistical significance). In other words, there was greater agreement on items that were judged to be at a lower depth of knowledge, and less agreement on items judged to be at a higher depth of knowledge.

Second, the level of rater agreement for an item was negatively correlated with item complexity (number of raters assigning a secondary topic rating to an item) for faculty raters but not for teacher raters. Among faculty, topic agreement was highest for items that were judged to be least complex (few secondary ratings were given) and was lowest for items that were judged to be most complex (many secondary ratings were given). The negative correlation is particularly noteworthy because the way the variables were coded was biased toward producing a positive correlation between depth of knowledge and agreement. Items for which raters gave both a primary and secondary rating had a higher probability of agreement than items for which raters gave only a primary rating by chance alone. That is, the former items provided two opportunities for a rater to agree with the majority opinion, rather than just one.

Table 15

Correlations Among Rater Topic Agreement, Depth of Knowledge, Item Complexity, and Item Difficulty Across Items

	Depth of Knowledge	Item Complexity	Item Difficulty
	Teachers and Faculty		
Rater topic agreement	-.32*	-.22	-.13
Average depth of knowledge		.50**	.37*
Item complexity (number of secondary ratings)			.12
Item difficulty			
	Faculty		
Rater topic agreement	-.26	-.39*	-.21
Average depth of knowledge		.51**	.38*
Item complexity (number of secondary ratings)			.19
Item difficulty			
	Teachers		
Rater topic agreement	-.19	.10	-.04
Average depth of knowledge		.54**	.34*
Item complexity (number of secondary ratings)			.07
Item difficulty			

Third, rater agreement did not correlate significantly with item difficulty as indicated by student performance.

Correlations among item features. The remaining correlations in the matrices in Table 15 concern the intercorrelations among various features of the items. First, depth of knowledge correlated positively with item complexity (whether items were judged to require both a primary and secondary topic rating), and the correlations were statistically significant for both faculty raters and for teacher raters. Items that raters judged to require high depth of knowledge were often the same items that raters judged to be most complex (many raters gave a secondary topic rating). Second, items that raters judged to require high depth of knowledge were often the same items that students found most difficult. Third, item complexity did not correlate significantly with item difficulty. Items on which many raters gave a second topic rating were not found by students to be more difficult than items on

which few raters gave a secondary rating. In summary, these findings show, first, that raters judged items that spanned two topics to require more depth of knowledge than items that required only one topic assignment, and, second, that students found items to be most difficult when they were judged to have high depth of knowledge but not when they spanned two topics.

Summary and Conclusion

Our study demonstrates one approach to examining the alignment of an assessment with a set of standards or expectations. The results show the feasibility of the process, and the process yielded reasonable psychometric results. Panels of high school educators and University faculty were able to engage in a structured rating process, and even though the statement of competencies on which the ratings were based was not designed for the purpose, raters achieved relatively high levels of agreement in the identification of specific topics and categories assessed by individual items. The majority of raters in both groups agreed on the content classification of the great majority of items, and kappa coefficients confirmed moderate to good agreement amongst faculty and teachers on topic and category assignments. The exceptions were topic areas where there were few items—statistics and probability—which only contained six items, and topic ratings for the two open-ended items. While the depth-of-knowledge scale appears to need some work to better differentiate various levels of intellectual demand, study results do provide important empirical verification of the ratings. There was a strong relationship between depth-of-knowledge ratings and student performance. It was heartening as well to see the agreement between how the developers of the test classified the test items and the classifications made by the educators in this study. Certainly, it will be important to extend work to other subject areas and other grade levels, but based on the results here, the technical underpinnings of the alignment process demonstrated here appear promising.

At the same time, our work raises questions and shows some of the challenges of aligning tests with standards and instruction. One important issue is the complexity of the construct itself and how to come to an overall judgment of the quality of alignment. Clearly one could reach different conclusions about the alignment between the test and the Statement on Competency depending on whether one used the criteria of relevance, comprehensiveness, or balance. Virtually all of the items were relevant to the standards, in terms of being centrally related to

at least one of the topics in the standards; the percentage of topics and categories "covered" showed relatively most attention to Algebra I, while the percentage of items aligned with each topic category told a different story. Analysis of both comprehensiveness and balance, however, did show that Probability and Data Analysis and Statistics got relatively little attention, and that these areas were assessed at more routine levels.

What is the optimal balance in these criteria? It seems reasonable to expect that all items on a test should be at least relevant to the standards that are intended to be objects of assessment, but in real practice this is not always the case. Certainly there needs to be some comprehensiveness in coverage, but depending on the purpose of the test, the issue is not only ticking off how many of the intended topics or standards are "covered" but whether the mix actually addresses the intended domain and whether there is sufficient coverage of each to permit desired inferences. For example, we know there is a big difference between a test that reflects or addresses multiple standards and those that permit diagnostic inference on student performance in each. For the former purpose, we need comprehensiveness and balance; for the latter, we need depth in the comprehensiveness, an issue that was not pursued here. Values as well as practical considerations come into the process. Whose values are/should be represented by the alignment of the test? To what extent are decisions made in advance and based on value decisions, as opposed to what items survive an empirical field test or happen to be on an off-the-shelf test? How comprehensive, deep, and balanced can an assessment be, given the reality of available testing time? These are some of the many questions that need to be addressed. Furthermore, such questions must be addressed early in the test design process so that specifications firmly aligned with test purpose(s) can be a solid basis for item and test development. Alignment considerations need to precede the test development or selection process, not trail it. Waiting for the results of an after-the-fact alignment study clearly is too late.

Our study too sidestepped the issue of how much agreement there really needed to be to assign an item to a particular standard. We somewhat arbitrarily used two common sense criteria—at least a majority (more than 50%) and a clear majority (70%). While the patterns of alignment in terms of comprehensiveness and balance were similar for both agreement levels, clearly the two criteria result in different substantive levels of alignment. Which criteria to use is a judgment call. From a feasibility standpoint, what level of agreement is reasonable to expect may

well vary with the complexity of the test target. Our findings suggest a relationship between levels of agreement and the cognitive demand and complexity of the test items. There was lower agreement on items that were judged to be at higher depth of knowledge, and there tended to be lower levels of agreement on items that were judged to assess more than one topic.

The challenge of reaching agreement on item classification may well extend to the problem of communicating and enabling teachers and students to understand what is expected of them. That is, it may be difficult for teachers who do not agree or think that a given standard translates into the kind of performance represented by specific items on the test to teach the standard in a way that is reflected in test performance. As our expectations for students become ever higher and the demand for standards that reflect complex thinking and problem solving become common, the problem of truly understanding what is expected and how to teach and learn it is likely to intensify. For example, if assessment items cross standards, competence for students means not only mastering the individual standards, but more complex performances that require these standards in novel combination.

Levels of agreement thus may represent important communication issues that need to be adequately addressed in standards-based reform. Is the glass half empty or half full? We were pleased with the agreement levels we were able to achieve with very limited training and orientation. Yet our findings still mean that sizeable proportions of educators did not agree on what our standards meant, at least in terms of their implications for testing. Furthermore, considering the expertise and experience of the educators and faculty who were involved in this study, as well the mathematics focus of the study, our findings may well represent a best case. The challenge of alignment continues.

References

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

California Department of Education (2001a, October). *High school mathematics blueprint,* (Draft). Sacramento, CA: California State Department of Education.

California Department of Education (2001b, October). *Overview specifications for high school mathematics* (draft). Sacramento, CA: California State Department of Education.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Committee on Equivalency and Linkage of Educational Tests, Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council). Washington, DC: National Academy Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378-382.

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).

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2000). *Benchmarking and alignment of standards and testing* (Draft Deliverable to OERI, Contract No. R305B960002). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

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.

Watkins, M. W., & Pacheco, M. (2001). Inter observer agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10,* 205-212.

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.

Appendix: Competency Topic List

Competency Topic List	
	Topics[a]
	Variables, Equations, and Algebraic Expressions
Essential for all entering college students	Algebraic symbols and expressions
Essential for all entering college students	Evaluation of expressions and formulas
Essential for all entering college students	Translation from words to symbols
Essential for all entering college students	Solutions of linear equations and inequalities
Essential for all entering college students	Absolute value
Essential for all entering college students	Powers and roots
Essential for all entering college students	Solutions of quadratic equations
Essential for all entering college students	Solving two linear equations in two unknowns including the graphical interpretation of a simultaneous solution
Essential for college students to be adequately prepared for quantitative majors	Solutions to systems of equations, and their geometrical interpretation
Essential for college students to be adequately prepared for quantitative majors	Solutions to quadratic equations, both algebraic and graphical
Essential for college students to be adequately prepared for quantitative majors	The correspondence between roots and factors of polynomials
Essential for college students to be adequately prepared for quantitative majors	The binomial theorem.
	<u>Families of Functions and Their Graphs</u>
Essential for all entering college students	Applications
Essential for all entering college students	Linear functions
Essential for all entering college students	Quadratic and power functions
Essential for all entering college students	Exponential functions
Essential for all entering college students	Roots
Essential for all entering college students	Operations on functions and the corresponding effects on their graphs
Essential for all entering college students	Interpretation of graphs
Essential for all entering college students	Function notation
Essential for college students to be adequately prepared for quantitative majors	Functions in context, as models for data
Essential for college students to be adequately prepared for quantitative majors	Logarithmic functions, their graphs, and applications
Essential for college students to be adequately prepared for quantitative majors	Trigonometric functions of real variables, their graphs, properties including periodicity, and applications
Essential for college students to be adequately prepared for quantitative majors	Basic trigonometric identities
Essential for college students to be adequately prepared for quantitative majors	Operations on functions, including addition, subtraction, multiplication, reciprocals, division, composition, and iteration
Essential for all entering college students	Polynomials
Essential for college students to be adequately prepared for quantitative majors	Inverse functions and their graphs
Essential for college students to be adequately prepared for quantitative majors	Domain and range.

Competency Topic List	
	Topics[a]
	Geometric concepts
Essential for all entering college students	Distances, areas, and volumes, and their relationship with dimension
Essential for all entering college students	Angle measurement
Essential for all entering college students	Similarity
Essential for all entering college students	Congruence
Essential for all entering college students	Lines, triangles, circles, and their properties
Essential for all entering college students	Symmetry
Essential for all entering college students	Pythagorean Theorem
Essential for all entering college students	Coordinate geometry in the plane, including distance between points, midpoint, equation Of a circle
Essential for all entering college students	Introduction to coordinate geometry in three dimensions
Essential for all entering college students	Right angle trigonometry
Desirable for all entering college students	Transformational geometry, including rotations, reflections, translations, and dilations
Desirable for all entering college students	Tessellations
Desirable for all entering college students	Solid geometry
Desirable for all entering college students	Three-dimensional coordinate geometry, including lines and planes
Essential for college students to be adequately prepared for quantitative majors	Two- and three-dimensional coordinate geometry
Essential for college students to be adequately prepared for quantitative majors	Locus problems
Essential for college students to be adequately prepared for quantitative majors	Polar coordinates
Essential for college students to be adequately prepared for quantitative majors	Vectors
Essential for college students to be adequately prepared for quantitative majors	Parametric representations of curves
	Probability
Essential for all entering college students	Counting (permutations and combinations, multiplication principle)
Essential for all entering college students	Sample spaces
Essential for all entering college students	Expected value
Essential for all entering college students	Conditional probability
Essential for all entering college students	Area representations of probability
	Data Analysis and Statistics
Essential for all entering college students	Presentation and analysis of data
Essential for all entering college students	Mean, median and standard deviation
Essential for all entering college students	Representative samples
Essential for all entering college students	Using lines to fit data and make predictions
Desirable for college students who intend to declare quantitative majors	Continuous distributions
Desirable for college students who intend to declare quantitative majors	Binomial distributions
Desirable for college students who intend to	Fitting data with curves

Competency Topic List	
	Topics[a]
declare quantitative majors	
Desirable for college students who intend to declare quantitative majors	Regression
Desirable for college students who intend to declare quantitative majors	Correlation
Desirable for college students who intend to declare quantitative majors	Sampling
	Argumentation and Proof
Essential for all entering college students	Mathematical implication
Essential for all entering college students	Hypotheses and conclusions
Essential for all entering college students	Direct and indirect reasoning
Essential for all entering college students	Inductive and deductive reasoning
Essential for college students to be adequately prepared for quantitative majors	Mathematical induction
Essential for college students to be adequately prepared for quantitative majors	Formal proof
	Discrete Mathematics
Desirable for all entering college students	Graph theory
Desirable for all entering college students	Coding theory
Desirable for all entering college students	Voting systems
Desirable for all entering college students	Game theory
Desirable for all entering college students	Decision theory
	Sequences and Series
Desirable for all entering college students	Geometric and arithmetic sequences and series
Desirable for all entering college students	The Fibonacci sequence
Desirable for all entering college students	Recursion relations
	Number Theory
Desirable for all entering college students	Prime numbers
Desirable for all entering college students	Prime factorization
Desirable for all entering college students	Rational and irrational numbers
Desirable for all entering college students	Triangular numbers
Desirable for all entering college students	Pascal's triangle
Desirable for all entering college students	Pythagorean triples
	Vectors and Matrices
Desirable for college students who intend to declare quantitative majors	Vectors in the plane
Desirable for college students who intend to declare quantitative majors	Complex numbers and their arithmetic
Desirable for college students who intend to declare quantitative majors	Vectors in space
Desirable for college students who intend to declare quantitative majors	Dot and cross product, matrix operations and applications

Competency Topic List	
	Topics[a]
	Conic Sections
Desirable for college students who intend to declare quantitative majors	Representations as plane sections of a cone
Desirable for college students who intend to declare quantitative majors	Focus-directrix properties
Desirable for college students who intend to declare quantitative majors	Reflective properties
	Non-Euclidean Geometry
Desirable for college students who intend to declare quantitative majors	History of the attempts to prove Euclid's parallel postulate
Desirable for college students who intend to declare quantitative majors	Equivalent forms of the parallel postulate
Desirable for college students who intend to declare quantitative majors	Models in a circle or sphere
Desirable for college students who intend to declare quantitative majors	Seven-point geometry.
	None of the above: Doesn't occur prior to Algebra 1
	None of the above: Occurs in curriculum prior to Algebra 1