

Chapter 9

Alignment Among Secondary and Post-Secondary Assessments in California

Vi-Nhuan Le, Laura Hamilton, and Abby Robyn
The RAND Corporation



Background

As students progress through high school and into institutions of higher education, they take numerous tests that vary in scope, content, and purpose. At the K-12 level, almost all of the states are currently using or developing assessments that are aligned with state standards.¹ Some of these assessment programs rely on commercially available, standardized, multiple-choice tests administered to every student, whereas others hire their own staff to develop items in multiple formats (including, for example, essays or portfolios) which are then administered in a matrix sampling scheme (i.e., not every student completes every item). In some states, scores on these tests are used to inform decisions about grade promotion and graduation. Students who plan to attend college also take one or more admissions tests, such as the ACT or the SAT I and II, and may take Advanced Placement (AP) exams, which provide college credit for high school coursework. When they arrive at college, many students are asked to take placement exams which are used to assign them to appropriate courses. These exams are especially prevalent in large state higher-education systems.

Assessments play a prominent role in the transition from high school to college. In most cases, test scores are among the major criteria used to determine who is accepted into an institution and who is assigned to remedial courses. Although these scores are imperfect, indirect measures of what students have accomplished, they often provide valuable information that may improve the decision-making process. A well-constructed test of achievement in a particular subject area constitutes a sample of performance from a larger domain to which the user wishes to generalize. This domain will vary depending in part on the purpose of the assessment. A statewide achievement test might be designed to sample from a range of topics and to cover material learned across several grades. A college placement exam, in contrast, may have a narrower focus, reflecting the curriculum of a particular course. Consequently, these tests may not resemble one another closely in the constructs that they measure. In other words, they may not be well aligned.

The goal of the present study is to investigate the degree of alignment among these different types of tests in six case-study states, and to explore the potential consequences of any misalignment. We will compare assessments

used for college admissions, college placement, and K-12 system monitoring and accountability in each state, classifying items along several dimensions. For each state, we will obtain a summary of the ways in which the assessments are and are not aligned with one another, and discuss possible implications. This report presents the results for California. It is important to note that we do not claim that all tests must be well aligned. The conditions under which alignment is important are discussed later.

This analysis is part of a larger study commissioned by Stanford University. “The Bridge Project: Strengthening K-16 Transition Policies” is a national study funded by the Pew Charitable Trusts and the U.S. Department of Education’s Office of Educational Research and Improvement. It focuses on the need to increase the alignment between higher education admissions-related requirements and K-12 curriculum frameworks, standards, and assessments. The study was prompted in part by a perceived disjuncture between standards for college admission and placement, on the one hand, and high school curriculum and instruction on the other.² The salience of this problem is underscored by a 1995 National Education Association survey in which 82 percent of House and Senate Education chairs polled viewed the improvement of connections between colleges and schools as among their highest priorities for higher education.³ Admissions policies are a primary way in which colleges influence the education of secondary students, and the tests that are given as part of the admissions and placement processes are a major component of these policies.

Importance of Alignment

There are at least three major ways to think about alignment among different assessments. First, the content and format of test items send messages to students who take them. Particularly when tests have high stakes attached, such as graduation from high school, selection into college, or placement into a remedial program, they can be expected to influence the behaviors of examinees and, in some cases, their instructors. For example, multiple-choice tests are often criticized for encouraging an emphasis on memorization of discrete facts rather than extended problem solving. It is important to determine whether tests are sending a consistent message to students regarding what kinds of knowledge and skills are valued by the institutions they wish to attend. It is also critical that students have ample opportunity to prepare in appropriate ways for high-stakes assessments. If students enter college unaware of what skills they will be expected to demonstrate on a placement exam, they may not perform as well as they would if given the opportunity to prepare. It is important to note here that the signals a test sends are somewhat distinct from the measurement properties of the test. For example, it is possible that a multiple-choice test does indeed measure complex problem-solving skill, but that examinees and instructors perceive the test as being focused on memorization or recall.

The importance of aligning the various aspects of the educational system to support a common set of goals has been recognized by advocates of systemic reform,⁴ promoters of test-based accountability systems,⁵ and many

others involved in educational reform efforts. Especially important to standards-based reform efforts is the degree to which the standards and the assessments used to measure progress toward them are consistent with one another. A recent study by Webb found varied degrees of alignment between tests and standards in math and science in four states.⁶ A content analysis of teacher licensing tests conducted by the Education Trust showed that most such tests required little more than high school level knowledge but that some were more rigorous than others.⁷ Standards and assessments that are not aligned with one another or that encourage a focus on low-level skills create mixed messages and confusion for students, teachers, and others involved in promoting student learning.

The second aspect of alignment involves the consistency with which students are rank ordered or classified into categories or programs (e.g., remedial instruction) by different tests. If two tests are designed to measure the same abilities, evidence must be gathered to show that students who do well on one tend to do well on the other. Although most tests of academic achievement tend to correlate highly with one another, even when subject and item format differ, it is nonetheless important to evaluate the magnitude of this correlation and the consistency of any classification that results from test use. Scores on a high school math exam should, for example, correlate highly with scores on a math placement test administered by the higher education system.

Finally, it is essential that the standards used for decision making be comparable across assessments and set in a technically sound and credible manner. The placement process often

involves selecting a cut score on an exam and assigning students to programs or courses based on whether or not their scores exceeded this cut score. Statewide assessment programs are increasingly reporting student performance in terms of standards similar to the achievement levels used on the National Assessment of Educational Progress (NAEP). These efforts have been criticized in part because the process of mapping performance to descriptors relies heavily on judgments that are often error-prone.⁸ Even so, assessment results continue to be reported in terms of standards, and it is therefore important to determine whether the standards set on different tests provide reasonably consistent information about students. If a student is labeled “Advanced” or “Proficient” on a state test but is unable to reach the level of performance on a placement test necessary to avoid remedial coursework, there is reason to believe that the standards used on one or both tests are inappropriate.

The current project is designed to provide information concerning the degree and nature of alignment among tests used for K-12 system monitoring and accountability, college admissions, and college placement in six states. The project is limited in scope and will not be able to address all forms of alignment. We will rely on expert judgments regarding the features that characterize test items, thereby addressing the first aspect of alignment discussed above. Because we will not have access to test score data, we will not be able to examine item characteristics or relationships among scores on different tests and criterion measures (such as first-year grade point average). A comprehensive study of standard-setting across instruments is also beyond the scope of this project.

Importance of Considering Purpose of Assessment

The degree of alignment among different sets of tests will undoubtedly vary substantially. Even when assessments are designed to be parallel, as with alternate forms of the SAT, we would not expect perfect alignment. Because the assessments we are comparing in this study were designed for different purposes, the alignment is likely to be much less than perfect. This is not necessarily a problem, if the differences result from appropriate efforts to tailor the measure to the situation for which it was designed. For example, a low-stakes K-12 system monitoring exam (i.e., one that is used to track achievement but that has no consequences for individual students, teachers, or schools), might be designed to include a broad variety of topics and therefore may not sample adequately from college-level material. There may be no discernible negative effect of this on students' efforts to prepare for other exams. If, however, scores on this K-12 exam were used to determine which students should graduate or which teachers should get bonuses in their paychecks, there would be a significant risk of "teaching to the test" that might result in teachers and students neglecting material that is not tested. This type of response has been observed in states with test-based accountability systems.⁹ Thus the purposes of the tests, and how they are viewed by school personnel and students, influence the degree to which misalignment may pose a problem.

The nature of the misalignment is also important. In the example presented above, the issue was primarily one of content sampling. The problem may be more serious when two tests reflect different philosophies concerning

what students should know and what kinds of skills they should be able to display. In many cases, the misalignment among K-12 and university-level tests results from reforms that have taken hold at one level of the educational system but not another. This is particularly true in states where new tests have been developed to reflect state standards or frameworks that emphasize inquiry-based teaching and open-ended problem solving. In such cases, the skills and knowledge students are expected to demonstrate on the state exams may differ substantially from what is expected on college admissions and placement exams. This creates a confusing set of signals for students concerning how they should prepare for the admissions and placement process. It is this signaling function of tests that is the primary motivation for this alignment study.

Finally, the examinee population for which the test was designed, and the ways in which scores are used, must be considered. Exams that are intended to make fine distinctions among high-ability students need to include a large number of difficult items and may include topics that are covered in advanced courses. Such items would be less appropriate for a test that is administered to the entire public school population. So it would be reasonable to expect some misalignments. All of the results we discuss below should be interpreted with this in mind. Later we provide further discussion of the importance of considering purpose.

California's Assessment Environment

The current policy environment with respect to standards and assessments in California is

described in the chapter of this volume by Venezia. Students in California high schools, particularly those who plan to attend college, take a number of tests that differ in format and purpose. Below we discuss each of the assessments that we examined in this study. We study only mathematics and English/language arts tests, though many of the assessment programs discussed below include tests in other subjects as well.

Several of the tests we examined, including the SAT I, SAT II, ACT, and AP exams, are used nationally to aid in college admissions decisions. The SAT I, a three-hour, mostly multiple-choice exam that measures general mathematical and verbal reasoning, is intended to help predict success in college. Evidence of its validity for this purpose typically focuses on correlations with freshman grade point average. The SAT II is a one-hour multiple-choice test that assesses in-depth knowledge of a particular subject, and is used by admissions officers as an additional measure with which to evaluate student subject-matter competence. The SAT II is used primarily at the more selective institutions and is taken by far fewer students than is the SAT I. For this study, we examined the following SAT II tests: Mathematics IC, Mathematics IIC, Literature, and Writing. The ACT is an approximately three-hour exam consisting entirely of multiple-choice items. Used as an alternative measure to the SAT I in evaluating applicants chances of success in college, it assesses achievement in several academic subjects, including science, reading, writing, and math. The AP tests are used to measure college-level achievement in several subjects, and to award academic credit to students who demonstrate college-level proficiency. We

examined two AP exams: Calculus AB and English Language and Composition.

Students are encouraged to take the ACT or SAT I within their junior or senior years, whereas the most optimal time to take the SAT II or AP exams is within months of completing a relevant course. Students are typically required to take either the SAT I or ACT, and, at certain schools, several SAT II exams as part of the admissions process. While the AP tests are not a requirement, admissions officers are likely to view students with AP experience as better-prepared and more competitive applicants.

In addition to the college entrance tests, California students encounter several other assessments during their high school years. As part of its Standardized Testing and Reporting (STAR) program, California currently requires public schools to administer the Stanford Achievement Test, Version 9 (Stanford 9) in grades 2 through 11, published by Harcourt Educational Measurement. Scores on this one-hour multiple-choice test are used to monitor student achievement in basic academic skills, and allow comparisons to be made to a national sample of students. In spring of 1999 a set of augmentation items was administered to supplement the Stanford 9. These included 35 language arts items and 35 math items, which were designed to assess progress toward the state-adopted content standards. In grades 8-10, the specific math items administered were determined by the math course in which the student was enrolled. The augmented portion of STAR is still evolving, and we were unable to obtain the actual items administered to students. Therefore these items are not included in our analysis. Results from the 1999 STAR

administration indicate that the augmented items were difficult for students. The governor has proposed tying merit-based college aid to performance on these items; this and other proposed high-stakes uses of STAR make it highly likely that both students and teachers will increasingly focus their efforts on this testing program.

Students also have the option of taking the Golden State Exams (GSE), which are voluntary tests allowing high schools students to earn special recognition when they graduate. The GSEs are 90-minute tests containing both multiple-choice and open-ended items. They are intended to assess student achievement relative to state-adopted content standards in particular subject areas. We included five of these tests in our study: High School Mathematics, First Year Algebra, Geometry, Reading/Literature, and Written Composition. Some of the GSE assessments are similar to end-of-course exams (e.g., Algebra or Geometry), and are best taken while the students are currently enrolled in the course.

Other GSEs are comprehensive tests that cover the content of several courses (e.g., Reading/Literature, Written Composition, and High School Mathematics). Students wishing to take these tests are advised to wait until their junior or senior year of high school.

Test	Materials Examined	Time Limit	Number of Items	Tools
ACT	Full sample form	60 minutes	60 MC	Calculator
AP Calculus AB	Full form, 1997 released exam	Two 90-minute sections	40 MC 6 Free response	Graphing calculator on last 15 MC items
California State University Entry Level Mathematics Placement Exam	Sample items	75 minutes	65 MC	Calculator
Golden State Exam (Algebra)	Sample items	Two separate 45-minute sessions	30 MC 2 OE	Calculator, Ruler
Golden State Exam (Geometry)	Sample items	Two separate 45-minute sessions	30 MC 2 OE	Calculator, Ruler

Table 1. Structural Characteristics of the Tests: Mathematics

Finally, examinees applying to any of the 31 colleges under the California State University (CSU) and the University of California (UC) systems may be required to take a placement exam in math and/or English. Many of the community colleges also administer placement

exams. These tests are used to determine whether admitted students possess entry-level math and English skills. CSU has placement tests for both math and English, whereas UC administers a system-wide test only for English. The CSU system requires its students

obtain a minimum achievement level on the SAT I, SAT II, or ACT in order to be exempted from taking a placement exam. UC requires a minimum achievement level on either the SAT II or AP exam. Students not meeting the minimum standards under the CSU guidelines must take a 75-minute multiple-choice math exam, and/or a 105-minute English test, which contains both multiple-choice and essay items. Examinees not meeting the UC standards for English are required to take a two-hour essay exam. The community colleges administer a range of exams; we include the Santa Barbara City College English exam in this analysis as an example. All students planning to enroll in an English course at the Santa Barbara City College must take the 85-minute College Tests for English Placement before registration. The test, consisting of both multiple-choice and essay items, is used to place students in an appropriate English course.

Tables 1 and 2 list these testing programs and the type of information we were able to obtain for this study. For most tests, we used a single form from a recent administration or a full-length, published sample test. In a few instances where full-length forms were

Purpose	Framework	Content as Specified in Testing Materials
Selection of students for higher education	High school mathematics curriculum	Prealgebra (23%), elementary algebra (17%), intermediate algebra (15%), coordinate geometry (15%), plane geometry (23%) and trigonometry (7%)
Provide opportunities for HS students to receive college credit and advanced course placement	AP Calculus Course Description	Calculus
Assess whether admitted students possess entry level math skills	<i>Statement on Competencies in Mathematics Expected of Entering College Students</i> reviewed by faculty from CA community Colleges, CSU, and UC systems	Algebra I and II (60%), geometry (20%), data interpretation, counting, probability, and statistics (20%)
Monitor student achievement toward state-approved content standards, provide special diploma	<i>Mathematics Content Standards for California Public Schools, Kindergarten Through Grade 12</i> adopted by the State Board of Education Standards	First-year algebra
Monitor student achievement toward state-approved content standards, provide special diploma	<i>Mathematics Content Standards for California Public Schools, Kindergarten Through Grade 12</i> adopted by the State Board of Education Standards	Geometry

unavailable, we used published sets of sample items. This was the case for the CSU placement tests and the GSEs. As mentioned earlier, we were also unable to obtain the STAR augmentation items, but instead looked at the STAR Test Blueprints provided by the California Department of Education. For the English/language arts (ELA) tests, the table specifies whether the test includes each of three possible types of items: reading, objective (e.g., multiple-choice) writing, and essay writing. When interpreting results, the reader needs to keep in mind that the percentages we report for the CSU and GSE exams are not necessarily the same percentages that would be obtained if we had examined an actual test form. They do, however, provide rough indicators of the emphasis placed on various topics in the materials that are used by students to prepare for the exams.

Methodology

The alignment analysis involved two major phases. In phase 1, we developed a framework

Test	Materials Examined	Time Limit	Number of Items	Tools
Golden State Exam (High School Mathematics)	Sample items	Two separate 45-minute sessions	30 MC 2 OE	Calculator, Ruler
SAT I	Full sample form	Two 30-minute sessions One 15-minute session	35 MC 15 QC 10 GR	Calculator
SAT II-Level IC	Full sample form	60 minutes	50 MC	Calculator
SAT II-Level IIC	Full sample form	60 minutes	50 MC	Calculator
Stanford 9	Full form	60 minutes	48 MC	Calculator, Ruler
Stanford 9 augmentation items	Test blueprints			Calculator, Ruler
Notes				
MC = multiple-choice				
OE = open-ended				
GR = grid-in				
QC = quantitative comparison				

Table 1 continued. Structural Characteristics of the Tests: Mathematics

of specifications for each subject. We examined several existing assessment frameworks, such as those used to develop the National Assessment of Educational Progress (NAEP), and combined

them to produce a set of specifications that addressed the range of topics and item types appearing on the tests included in this study.

We then applied these frameworks to our set of tests, and made several rounds of modifications in response to difficulties we encountered in conducting the alignment. The process was similar to one that we use for developing scoring rubrics for open-ended assessment items. The resulting frameworks are described later in this report.

Phase 2 consisted of the actual alignment exercise. Two raters who had expertise in both the relevant subject area and in the application of scoring criteria to assessment results conducted the alignment analysis for each subject. The raters worked through several of the assessments together. When raters differed in their interpretations of the framework components, they discussed the difference until agreement was reached. In cases where a disagreement could not be resolved, a third rater determined the final categorization. This process resulted in reasonably high levels of agreement (kappa values of approximately 85 percent to 100 percent) for most categories. Two exceptions were content area in math, where items often assessed skills in more than one area, and passage topic in reading, because passages often could be coded as addressing more than one topic. A final exception was the cognitive process category in math, discussed further below. For

Purpose	Framework	Content as Specified in Testing Materials
Monitor student achievement toward state-approved content standards, provide special diploma	<i>Mathematics Content Standards for California Public Schools, Kindergarten Through Grade 12</i> adopted by the State Board of Education Standards	Algebra I and II, geometry, probability and statistics
Selection of students for higher education	High school mathematics curriculum	Arithmetic (13%), algebra (35%), geometry, (26%), and other (26%)
Selection of students for higher education	Three-year college preparatory mathematics curriculum	Algebra (30%), geometry (38%, specifically plane Euclidean (20%), coordinate (12%), and three-dimensional (6%)), trigonometry (8%), functions (12%), statistics and probability (6%), and miscellaneous (6%)
Selection of students for higher education	More than three years of college preparatory mathematics curriculum	Algebra (18%), geometry (20%, specifically coordinate (12%) and three-dimensional (8%)), trigonometry (20%), functions (24%), statistics and probability (6%), and miscellaneous (12%)
Monitor student achievement toward CA standards	<i>National Council of Teachers of Mathematics Standards</i>	Two subtests: mathematical problem-solving and mathematical procedures
Monitor student achievement toward CA standards	CA standards	23% algebra I, 31% geometry, 31% algebra II, 14% statistics

these categories, agreement tended to be approximately 70 percent.

Results for Mathematics

In this section we describe the results of the alignment exercise for math tests. First we present the framework that was developed. We then describe the major areas of alignment and misalignment, and discuss the implications of these findings for the signals that students receive.

Framework

The math framework consisted of three major dimensions: technical features, content, and cognitive processes. This set of dimensions was used in an earlier study of the alignment between state tests and NAEP,¹⁰ but we modified the definitions of these dimensions to some degree to reflect unique characteristics of some of the tests we examined in this study. The technical dimension covered features of the test that could be described through simple examina-

tion of the test and items—number of items, time limit, format (e.g., multiple-choice, essay), provisions for the use of tools such as calculators or protractors, the use of diagrams or other graphics, the use of formulas, and whether each item was embedded in a context (as in a word problem). The use of formulas

Test	Materials Examined	Time Limit	Number of Items	Purpose
ACT	Full sample form	80 minutes (35 minute reading section, 45 minute objective writing section)	40 MC reading 75 MC objective writing	Selection of students for higher education
AP Language and Composition	Sample questions	60 minute MC section 120 minute essay section	52 MC 3 essays	Provide opportunities for HS students to receive college credit and advanced course placement
California State University Entry Level English Placement Exam	Sample items	Two 30-minute sections (one section each for reading and objective writing) 45 minute essay section	45 MC reading 45 MC objective writing 1 essay	Assess whether admitted students possess entry level English skills
Golden State Exam (Reading/Literature)	Sample items	Two separate 45-minute sessions	30 MC 2 essays	Monitor student achievement toward state-approved content standards, provide special diploma
Golden State Exam (Written Composition)	Sample Items	Two separate 45-minute sessions	30 MC 2 essays	Monitor student achievement toward state-approved content standards provide special diploma

Table 2. Structural Characteristics of the Tests: English/Language Arts

was sometimes difficult to determine because problems can be solved in multiple ways, and in some cases an item could be solved either with or without a formula. Items were coded as requiring a formula only if it was determined that the formula was necessary for solving the problem. Finally, we examined the context sur-

rounding the assessment, particularly the degree to which high stakes are attached to performance. This is important because it affects examinee motivation.

The content dimension included several categories of math topics, from pre-algebra (e.g., numbers and operations) through calculus.

Almost all of the tests we examined had specifications that included many or all of these categories. We listed sub-categories as a means of making the distinctions among the main categories clearer, but we coded using only the main categories.

Finally, the cognitive dimension was identical to that used for NAEP, and included three categories—conceptual understanding, procedural knowledge, and problem solving. As is typical with studies like this, the raters found this dimension to be the most difficult to code.¹¹ The cognitive process categories cannot always be separated neatly: According to the NAEP framework, “These abilities are...descriptions of the ways in which information is structured for instruction and the ways in which students manipulate, reason with, or communicate their mathematical ideas. As a consequence, there can be no singular or unanimous agreement among educators about what constitutes a conceptual, a procedural, or a problem-solving item. What can be classified are the actions a student is likely to undertake in processing information and providing a satisfactory response.”¹²

Framework	Reading Section?	Objective Writing Section?	Essay Section?
High school mathematics curriculum	Y	Y	N
AP English Language and Composition Course Description	Y	N	Y
CSU English curriculum	Y	Y	Y
<i>English-Language Arts Content Standards for California Public Schools, Kindergarten Through Grade Twelve</i> , adopted by the State Board of Education Standards	Y	N	Y
<i>English-Language Arts Content Standards for California Public Schools, Kindergarten Through Grade Twelve</i> , adopted by the State Board of Education Standards	N	Y	Y

In addition, items can often be solved in multiple ways, sometimes as a function of the examinees proficiency. What might be a problem-solving item for one examinee might require another to apply extensive procedural knowledge. For instance, consider an item asking students for the sum of the first 101 numbers starting with zero. A procedural knowledge approach might involve a computation-intensive method, such as entering all the numbers into a calculator to obtain the resulting sum. However, the problem-solving approach would entail a recognition that all the numbers, except the number 50, can be paired with another number to form a sum of 100 (100+0, 99+1, 98+2, etc.). The total sum is then simply computed by multiplying the number of pairs (i.e., 50) by 100 and adding 50. Clearly, depending upon the chosen approach, the same item can elicit varying levels of mathematical sophistication. The cognitive processes required by the items affect the construct that they measure and, as a consequence, examinee scores. However, for the purposes of this study, which focuses on signals sent to examinees, clear distinctions along this dimension are arguably less critical.

Aspects of Alignment and Misalignment in Mathematics

To evaluate alignment, the degree of consistency among the measures in connection with structural and content characteristics was studied.ⁱ Table 3 provides more details on the structural and content features of each test. The measures shared some features, particularly those related to format and administrative conditions. Every assessment included multiple-choice items, and all but the GSE were administered in a single testing session that took

Test	Materials Examined	Time Limit	Number of Items	Purpose
Santa Barbara College Tests for English Placement	Full sample form	85 minutes (30 minutes reading section, 35 minutes objective writing section, 20 minute essay)	35 MC reading 70 MC objective writing 1 essay	Assess whether students possess entry level English skills
SAT I	Full sample form	Two 30-minute sessions One 15-minute session	78 MC	Selection of students for higher education
SAT II-Literature	Full sample form	60 minutes	60 MC	Selection of students for higher education
SAT II-Writing	Full sample form	One 40-minute MC session One 20-minute essay session	60 MC 1 essay	Selection of students for higher education
Stanford-9	Full form	60 minutes	84 MC (54 reading comprehension items, 30 vocabulary items)	Monitor student achievement toward CA standards
University of California Subject A Examination	Sample questions	2 hours	1 essay	Assess admitted students' writing skills

Table 2 continued. Structural Characteristics of the Tests: English/Language Arts

approximately one hour. Students were allowed the use of a calculator, although most questions did not require extensive computation.

Familiarity with basic formulas and mathematical identities was generally assumed as background for the questions, but knowledge of more complex formulas was seldom necessary.

The assessments, however, tended to have many more differences than similarities. There was a great deal of structural variation among the exams, especially with regard to the percentages of items containing formulas and illus-

trations. Fewer than 10 percent of the SAT I and Stanford 9 items required a memorized formula, in contrast to 25 percent of the GSE Geometry problems. Whereas the GSE Algebra and SAT II Level IIC assessments made little use of figures, the Stanford 9 and GSE Geometry exams included many illustrations, with 42 percent and 75 percent of their items, respectively, containing a diagram.

Differences in the degree to which tests require interpretation of spatial or figural information are particularly important as they can affect gender and other group differences.

Instances of misalignment were also observed with respect to the amount of contextualization provided. In spite of reform ideology that recommends the inclusion of personally relevant items that require applications of mathematical principles to real-life situations, many of the exams continued to measure student achievement with abstract questions—that is, questions that included only numbers and symbols. No more than 25 percent of the items found on the college admissions and placement assessments were contextualized (i.e., embedded in a story), whereas more than half of the Stanford 9 items were classified as being contextualized.

Perhaps more important than the percent of contextualized items is the nature of the contextualization. In this respect, only the GSE open-ended questions were in line with the reform movement. Although 58 percent of the Stanford 9 items were framed in realistic situations, the presented sce-

Framework	Reading Section?	Objective Writing Section?	Essay Section?
High school Reading and Language Arts Curriculum	Y	Y	Y
High school Reading and Language Arts Curriculum	Y	Y	N
High school English and American Literature Curriculum	Y	N	N
High school Reading and Language Arts Curriculum	N	Y	Y
Aligned with NAEP framework	Y	N	N
UC English curriculum	N	N	Y

narios were brief, and had limited practical applications. On the other hand, the GSE open-ended items allowed examinees to impose their own meanings and constraints, and bore some relevancy to “real-world” skills. The GSE open-ended items will be discussed more fully in a later section.

Widespread reform efforts have also been directed toward the format in which test items are presented. Despite frequent criticisms that multiple-choice items are limited in the skills they measure, only the SAT I and GSE included items that required students to generate their own responses. The GSE open-ended questions, however, were much more extensive than the SAT I items. Successful solution of a GSE open-ended problem generally required multiple steps, and students were asked to justify or explain their solutions—frequently with diagrams or charts. In contrast, the SAT I open-ended items did not necessarily call for multiple strategies, and could sometimes be solved with algorithmic procedures. Furthermore, the SAT open-ended items were constrained, as the responses could not take on negative values. Thus, although the two tests

both make use of an open-response format, the cognitive demands differ dramatically.

An analogous problem arises with similarly named tests that assess very different sets of skills. Although all the exams are considered measures of mathematics achievement, there is a great deal of variation in the constructs assessed. Approximately 52 percent of the GSE Algebra items and 37 percent of the SAT I questions measured elementary

Test	Format				Context C	Graphs			Diagrams		
	MC	QC	GR	OE		S	RO	P	S	RO	P
ACT	100	0	0	0	22	5	2	0	13	0	0
CSU	100	0	0	0	24	0	0	0	16	0	0
GSE (Algebra)	95	0	0	5	15	0	5	0	10	0	0
GSE (Geometry)	95	0	0	5	10	0	0	5	75	0	0
GSE (HS Math)	92	0	0	8	33	0	5	0	23	0	5
SAT I	58	25	17	0	25	7	0	0	18	0	0
SAT II-Level IC	100	0	0	0	18	8	0	0	26	0	0
SAT II-Level IIC	100	0	0	0	12	12	2	0	2	0	0
Stanford 9	100	0	0	0	58	21	4	0	42	0	0

Legend:

Format MC = multiple-choice items QC = quantitative comparison items GR = fill-in-the-grid items OE = open-ended items	Context C = contextualized items RO = graph/diagram within response options P = graph/diagram needs to be produced
Formulas M = formula needs to be memorized G = formula is provided	Content PA = prealgebra EA = elementary algebra IA = intermediate algebra CG = coordinate geometry PG = plane geometry TR = trigonometry SP = statistics and probability MISC = miscellaneous topics

Table 3. Percent of Items Falling in each Category: Mathematics

algebra knowledge, whereas 40 percent of the Stanford 9 items focused on statistics. For college admissions exams such as the SAT II Level IIC and ACT, relatively greater emphasis was given to trigonometry, a topic that was absent from the both the GSE Algebra and SAT I exams.

The misalignments among the measures go beyond content sampling, and extend to the

reasoning requirements elicited by each test. Although none of the assessments focused heavily on problem-solving items, there were some differences with respect to the emphasis given to domain knowledge. Ninety-eight percent of the CSU items entailed straightforward application of declarative and procedural knowledge. In a similar vein, the vast majority of questions on the ACT, Stanford 9, and SAT

II Level IC tests were also solvable via heuristics and algorithms. The SAT II Level IIC, which was intended for examinees enrolled in more advanced college preparatory math courses, placed the most emphasis on problem-solving ability (20 percent of its questions).

Perhaps the source of the inconsistencies can be traced to variations in the purposes of the assessments and in the frameworks that guided their development. The GSE and CSU were designed to be aligned with state-adopted content standards, which have clearly prescribed guidelines that shape the content of the assessments. The Stanford 9 also employs an external framework, the *National Council on Teachers of Mathematics Standards*, but this set of guidelines encompasses standards that cut across state lines. Because they do not follow any explicit framework, the college admissions exams that assess knowledge in particular subjects (i.e., SAT II and ACT) have more loosely defined standards, and draw upon core concepts taught within most mathematics courses. The SAT I, on the

Formulas		Content								Cognitive Requirements		
M	G	PA	EA	IA	CG	PG	TR	SP	MISC	CU	PK	PS
15	0	17	22	5	15	25	8	3	5	40	53	7
18	0	6	32	8	16	14	2	22	0	28	70	2
10	0	0	52	0	19	14	0	10	5	19	76	5
25	0	0	0	0	5	86	10	0	0	52	38	10
15	0	23	15	0	23	23	0	15	0	62	23	15
1	8	13	37	2	6	19	0	13	11	32	53	15
12	0	2	30	10	12	28	4	8	6	34	58	8
10	0	2	14	22	12	14	18	6	12	26	54	20
6	6	0	13	2	19	19	4	40	4	63	31	6

Graphs/Diagrams
S = graph/diagram within item-stem

Cognitive Requirements
CU = conceptual understanding
PK = procedural knowledge
PS = problem-solving

other hand, is independent of any specific curriculum or course, and is intended to assess general mathematical reasoning proficiency developed over years of schooling.

Several of the misalignments discussed earlier should probably not be considered problematic, as some of the differences emerge from appropriate efforts to adapt a test to serve a particular purpose. For instance, although both the SAT Level IIC and Stanford 9 included topics from a wide variety of courses, the SAT Level IIC drew upon trigonometry, whereas the Stanford 9 rarely included such material. The broad content sampling found on both of these assessments can be further contrasted with the topics on the GSE Geometry test, which reflected the curriculum of a specific course. In this particular case, the Stanford 9, SAT Level IIC, and GSE Geometry exam have disparate purposes, which call for differing levels of mathematical sophistication and varying extent of domain sampling. They are also targeted toward somewhat different examinee populations. Because the SAT Level IIC is typically used to select among higher-achieving students for entrance into universities and colleges, the test needs to include many complex problems with advanced content in order to distinguish among the examinees and rank order them consistently. The Stanford 9, on the other hand, is used to monitor K-12 student achievement, and therefore require items of more moderate difficulty that can be attempted by students with a wider range of proficiency levels and course-taking histories. In a similar vein, the GSE Geometry test, unlike the SAT Level IIC or the Stanford 9, is not a measure of general math ability, but a measure of achievement in a particular course. Consequently, it is

more appropriate for this assessment to limit its content to a narrow area of math than to sample extensively from the entire mathematics domain. Thus, when making decisions concerning whether misalignments pose a potential problem, it is important to consider the use of the test. For the measures discussed above, the discrepancies most likely arise from variations in their purposes, and are therefore acceptable instances of misalignment.

However, discrepancies among exams with similar purposes are also evident. Consider the SAT II Level IC and GSE High School Math exams.ⁱⁱ Although both are intended to assess the proficiency of students who have taken three years of college preparatory math courses, they differ in their structural and cognitive features. The GSE contained a higher proportion of contextualized items (33 percent compared to 18 percent), whereas the SAT II included more graphs (8 percent compared to none). The GSE High School Math test also placed a greater emphasis on problem-solving items. Finally, there were vast differences in content sampling; the GSE was more likely to draw upon pre-algebra (23 percent compared to 2 percent), whereas the SAT II included more elementary algebra items (30 percent compared to 15 percent). In this particular case, the inconsistencies among the two sets of testing materials may send mixed messages to students regarding the emphases placed on various topics and skills.

Implications of the Misalignments

The misalignments among the exam materials can create a confusing set of signals pertaining to how students should prepare for the assessments. For example, the ACT and SAT I are