

often used interchangeably for college admissions, yet require students to demonstrate substantially different skills and knowledge. The ACT requires examinees to memorize formulas and identities, and includes numerous textbook-like problems that can be solved via simple application of procedural and declarative knowledge. In contrast, the SAT I provides students with formulas and mathematical identities, and places relatively more emphasis on abstract reasoning. Ostensibly, students preparing for the SAT I should not spend their time memorizing formulas, and should instead focus their efforts on furthering their inferential reasoning skills. On the other hand, examinees studying for the ACT might attempt to review formulas or practice algorithmic problems. Particularly for high-stakes exams, it is crucial that students prepare in appropriate ways, as differences in preparation efforts can greatly influence performance.

Inconsistencies among the exams are not the only potential source of confusion, as discrepancies between a framework and a test can send contradictory messages. *The Mathematics Content Standards for California Public Schools, Kindergarten Through Grade 12*, the National Council on Teachers of Mathematics (NCTM) Standards,¹³ and the *Statement on Competencies in Mathematics Expected of Entering College Students* indicate that desired outcomes of mathematics instruction include an increase in mathematical reasoning and communication, as well as a greater appreciation for the role that mathematics plays in everyday life. Of the three exams that resulted from these frameworks (the GSEs, Stanford 9 and CSU, respectively), the content of two of the tests did not appear to address these particular outcomes. Fewer than

10 percent of the items on the Stanford 9 and the CSU assessed problem-solving ability, and none required students to communicate mathematically. Furthermore, the plethora of abstract questions on the CSU exam, and the limited practical applications of the Stanford 9 contextualized items, may suggest to students that mathematics is not useful or relevant to real-world problems. The multiple-choice format favored by the Stanford 9 and CSU can also send negative messages regarding the importance of reasoning skills. Although items in any format can be designed to measure a variety of abilities, multiple-choice items are popularly believed to be less adequate than free-response questions at measuring higher-order thinking. Additionally, multiple-choice items are solution-oriented, as students who select the correct option receive full credit, regardless of the logic or reasoning underlying the given response.

The signals stemming from the Stanford 9 or CSU can be contrasted with those from the GSE. The GSE open-ended items were well-contextualized and process-oriented. The latter factor was clearly evident in the scoring rubric, which awarded different scores to two students who had the same set of calculations but who varied in their justifications of their work. In essence, scores were strongly affected by the degree to which students communicated their responses. However, the GSE test instructions were vague as to how elaborate the students' explanations should be, and in some instances the failure to receive the maximum number of points might have stemmed from a mismatch between the item stem and the scoring guidelines. For instance, one item presented students with data relating the amount of compression with the height of a ball shot upwards, and

asked students to “make a graph of this information.” Students choosing a bar graph received only partial credit because the bar graph was not the most appropriate manner in which to represent the data. Perhaps if the instructions were more specific in their requirements and prompted students to consider the most suitable manner of data representation (as opposed to any mode of representation), these students might have chosen a different type of graph. Especially for free-response items, the standards that will be used for judging responses must be clearly and adequately conveyed to examinees.

Results for English/Language Arts (ELA)

In this section we present the results of our analysis of alignment among tests used to assess students’ skills in reading and writing. The tests’ names were varied, but they all focused on reading and/or writing in the English language. Table 2, discussed briefly above, lists the tests along with basic details.

Framework

The ELA framework covers three types of items: reading, objective writing (mainly multiple-choice items), and essay writing. Many of the tests we examined included two or all three of these item types, whereas others focused on a single type. In contrast to mathematics, there were no clear content areas that could be used to categorize items. Instead, the ELA analysis focuses more on structural characteristics and cognitive demands. In addition, many of the tests include short passages followed by sets of

items, so it was necessary to categorize both the passage and the individual item.

There was extensive overlap among the frameworks for reading, objective writing, and essay writing. As with math, we identified subcategories to sharpen the distinctions among the main categories, but we coded using only the main categories. The structural dimensions, described in further detail in Table 4a, included three categories. The topic category captured the subject matter of the passage, and consisted of five areas—fiction, humanities, natural science, social science, and personal accounts. The type category identified the author’s writing style as narrative, descriptive, persuasive, or informative. The stimulus category referred to the presentation of the passage, such as a letter, essay, poem, or story. Raters used all three categories when coding the reading and objective writing items, but used only the topic category when coding the essay writing questions.

The cognitive framework for both the reading and objective writing measures consisted of a single cognitive demand dimension. Raters coded questions as assessing ability to recall information, make inferences, or evaluate an item’s style. In reading, questions that could be answered via direct reference to the passage were coded as recall items, whereas questions that required the examinees to interpret the material were coded as inference items.

Questions that pertained to the development of ideas or improved upon the presentation of the reading passages were coded as evaluating style.

For the objective writing measures, items that entailed application of grammatical rules were considered recall items. Typically, most of these questions concerned mechanics or usage errors.

Description or Example	
Type of Writing	
Narrative	Stories, personal essays, personal anecdotes
Descriptive	Describes person, place, or thing
Persuasive	Attempt to influence others to take some action or to influence someone’s attitudes or ideas
Informative	Share knowledge; convey messages, provide information on a topic, instructions for performing a task
Topic	
Fiction	story, poem
Humanities	e.g., artwork of Vincent Van Gogh
Natural sciences	e.g., the reproductive process of fish
Social sciences	e.g., one man, one vote; cost effectiveness of heart transplants
Personal	e.g., diary account of death of a parent
Stimulus materials	
Letters	
Essays	
Poems	
Stories	

Table 4a. Description of the ELA Structural Dimension Coding Scheme

Inference items were those that required examinees to identify cause-and-effect relationships, and “evaluating style” items asked students to display rhetorical skills including an understanding of sentence organization, clarity, and other stylistic features of written work. Table 4b gives more details of the cognitive coding systems.

The above framework was not applicable to the essay writing items, since all of the essay tests prompted students to establish and support a thesis. Students could use recalled knowledge as well as make inferences, and were asked to construct a clear presentation (see Table 4b). For the essay writing questions, raters focused on the scoring criteria, which

highlight the emphasis given to mechanics, word choice, organization, style, and insight.

Aspects of Alignment and Misalignment in English/Language Arts

We analyzed the degree of alignment among the different assessments by comparing the structural and content dimensions for each passage and each item. All of the ELA exams with reading sections used a passage as an item prompt, and virtually all of the studied tests included a set of multiple-choice items (the UC placement test was the exception).

Perhaps indicative of the loosely defined nature of the subject matter, differences among

the exams were much more prevalent than in math. Some assessments did not involve a written composition (ACT, SAT I, SAT II Literature, and Stanford 9) whereas others required two or three essays (AP, GSE Reading/Literature, and GSE Written Composition). There were also vast differences in the amount of time students were permitted to write their essays; the UC system allotted two hours for a single essay, whereas the SAT II Writing exam and the Santa Barbara City College exam each allowed only 20 minutes for essay completion.

The differences were not limited to the administrative characteristics of each exam, but were also apparent with respect to the structural features. In reading, all of the passages on the SAT II Literature test were narrative, and 63 percent were on fictional literary topics (see

Table 5a). In contrast, the SAT I passages tended to be informative (60 percent), and were much more likely to draw from humanities (40 percent). The essay was the most predominant presentation mode, with all of the passages on the AP, CSU, GSE Reading/Literature, and Santa Barbara City College exams presented in this manner. The majority of the passages on the SAT I and ACT were also essays (80 percent and 75 percent, respectively), but the SAT II Literature and the Stanford 9 varied the stimuli in which the reading passages were presented. The Stanford 9 included a letter and a flyer, whereas the SAT II Literature test was the only reading exam that included poems as a stimulus. These formats were not found on the other reading exams.

On measures of objective writing, the ACT, CSU, GSE Reading/Literature, SAT II

Description or Example		Used for Reading	Used for Object Writing	Used for Essay Writing
Cognitive Demands				
Recall	Answer can be found directly in the text, or by using the definitions of words or literary devices, or by applying grammatical rules	X	X	
Infer	Interpret what is already written	X	X	
Evaluate style	Improve the way the material is written	X	X	
Scoring Criteria				X
Mechanics	Grammar, punctuation, capitalization			X
Word choice	Use of language, vocabulary, sentence structure			X
Organization	Logical presentation, development of ideas, use of appropriate supporting examples			X
Style	Voice, attention to audience			X
Insight	Analytic proficiency, accurate understanding of stimulus passage			X

Table 4b. Description of the ELA Cognitive Dimension Coding Scheme

Writing, and Santa Barbara City College assessments included passages as item prompts, whereas the SAT I did not (see Table 5b). Virtually all the passages were presented as essays, although the Santa Barbara City College exam did include stories as a stimulus. There was some variation in the types of passages, as the GSE Written Composition passages were narrative, whereas the CSU passages were informative. Passages on the ACT, SAT II Writing, and Santa Barbara City College exams were approximately equally divided between narrative and informative. In a similar manner, the topics of the objective writing passages varied greatly from one test to the next; the ACT and SAT II Writing items tended to include themes from humanities (60 percent and 100 percent, respectively) whereas the CSU test focused on issues in social science. In contrast, the GSE Written Composition included personal accounts.

For the extended essay writing assessments, all the measures but the CSU and Santa Barbara City College exams included a topic from humanities (see Table 5c). Personal accounts were also commonly chosen prompts, found on such assessments as the AP, GSE Written Composition, GSE Reading/Literature, UC Subject A, and Santa Barbara City College tests. Of the forms that we studied, only the GSE Written Composition and CSU exams selected a social science theme, and only the UC Subject A test included a topic from natural science. None of the prompts drew from fictional material.

Inconsistencies among the exams were particularly evident with respect to the cognitive demands of each test. Of the reading assessments, only the AP test required students to

analyze a literary excerpt via a written composition. The remaining exams assessed knowledge and understanding of a passage solely with multiple-choice items. The cognitive complexity of the multiple-choice questions varied greatly among each of the measures. In reading, for instance, the SAT I and SAT II Literature tests placed great emphasis on analytical ability, with 83 percent and 80 percent of their items, respectively, assessing inferential skills (see Table 5d). Tests such as the AP and CSU also emphasized inferential skills, although not as heavily as the SAT I or SAT II Literature exams (77 percent and 66 percent, respectively). In contrast, measures such as the ACT, Stanford 9, and GSE Reading/Literature focused on straightforward recollection of information (58 percent, 71 percent, and 86 percent of their questions, respectively).

There was also great variation in cognitive complexity on the objective writing assessments (see Table 5e). Of the six measures, only the SAT I and the Santa Barbara City College included a significant proportion of items assessing inferential skills (100 percent and 57 percent of their items, respectively). Such questions comprised less than 5 percent of the items on the ACT and SAT II Writing exams, and were completely absent from the GSE Written Composition test. The CSU focused on evaluating writing style (64 percent), whereas GSE emphasized recall items (67 percent). Tests such as the ACT and SAT II Writing exams were more balanced in the kinds of skills they assessed; the items on these tests were mainly divided among recollection of information and evaluation of style.

There was much more consistency with respect to the kinds of cognitive demands

Test	Type						Topic		
	Narrative	Descriptive	Persuasive	Informative	Fiction	Humanities	Natural Science	Social Science	Personal Science
ACT	50	0	0	50	25	25	25	25	0
AP	75	0	0	25	0	25	25	0	50
CSU	100	0	0	0	0	100	0	0	0
GSE Reading/Literature	100	0	0	0	0	0	0	0	100
Santa Barbara City College	0	0	0	100	0	43	43	14	0
SAT I	40	0	0	60	20	40	20	20	0
SAT II Literature	100	0	0	0	63	0	0	13	25
Stanford 9	50	0	17	33	17	33	33	0	17

Table 5a. Percent of Reading Passages Falling into Each Category

Test	Type						Topic		
	Narrative	Descriptive	Persuasive	Informative	Fiction	Humanities	Natural Science	Social Science	Personal
ACT	40	0	0	60	0	60	20	0	20
CSU	0	0	0	100	0	0	0	100	0
GSE Written Composition	100	0	0	0	0	0	0	0	100
Santa Barbara City College	50	0	0	50	0	100	0	0	0
SAT I	0	0	0	0	0	0	0	0	0
SAT II Writing	50	0	0	50	0	100	0	0	0

Table 5b. Percent of Objective Writing Passages Falling into Each Category

Test	Topic				
	Fiction	Humanities	Natural Science	Social Science	Personal Essay
AP		X			X
CSU				X	
GSE Reading/Literature		X			X
GSE Written Composition		X		X	X
SAT II Writing		X			
Santa Barbara City College					X

Table 5c. Topic Contents of Essay Writing Prompts

	Stimulus			
	Letter	Essay	Poem	Story
	0	75	0	25
	0	100	0	0
	0	100	0	0
	0	100	0	0
	0	100	0	0
	0	80	0	20
	13	25	50	13
	17	33	0	50

	Stimulus			
	Letter	Essay	Poem	Story
	0	100	0	0
	0	100	0	0
	0	100	0	0
	0	50	0	50
	0	0	0	0
	0	100	0	0

required by measures of writing ability (see Table 5f). Skills such as mechanics, word choice, style, organization, and insight were identified as important factors in virtually all of the tests we studied. However, the GSE Reading/Literature test downplayed the importance of mechanics, word choice, and style, and the SAT II Writing test did not identify insight as part of its scoring criteria. The implications of these omissions will be discussed later.

As was the case with math, two verbal tests may have the same construct label, yet make vastly different cognitive demands. The GSE Reading/Literature, AP Literature and Composition, and SAT II Literature test are all

measures of reading proficiency, but differ in the kinds of skills assessed. The GSE Reading/Literature items typically entailed recollection of facts directly from a given passage, and usually did not ask students to judge the mood or tone of the piece. Both the AP and SAT II Literature assessments, on the other hand, required deeper analysis of the reading passage, oftentimes asking students to determine the effect of a given line or infer the intentions of the author. The AP exam, in particular, required students to apply their knowledge of literary devices. The AP test included many items asking students to identify examples of hyperboles, alliterations, and the like, but such questions were not found on either the GSE Reading/Literature or the SAT II Literature exams.

Discrepancies between the curricular standards and the tests were also apparent. For instance, the ability to learn the meaning of a word from context is perceived to be an integral aspect of English, yet most of the tests did not address this skill. Instead, many of the vocabulary items assessed students’ recall ability rather than their inferential skills. The ACT, AP, GSE Reading/Literature, SAT II Literature, and Stanford 9 assessments typically framed a vocabulary item as follows: “In lines XX, the word ‘panacea’ is best understood to mean...”. Although the question is phrased to indicate that the meaning relies on context, it can be construed as a recall question, as *a priori* knowledge of the definition is sufficient for a correct answer, since the context of lines XX did not affect the standard definition of “panacea.”ⁱⁱⁱ

Two tests that did ask examinees to derive meaning from context were the CSU and the SAT I. The CSU contained a section in which

Test	Recall	Infer	Evaluate Style
ACT	58	42	3
AP	23	77	0
CSU	33	66	0
GSE Reading/Literature	86	14	0
Santa Barbara City College	54	46	0
SAT I	18	83	0
SAT II Literature	13	80	7
Stanford 9	71	29	0

Table 5d. Percent of Reading Items Falling into Each Category

Test	Recall	Infer	Evaluate Style
ACT	48	4	48
CSU	14	21	64
GSE Written Composition	67	0	33
Santa Barbara City College	16	57	27
SAT I	0	100	0
SAT II Writing	50	3	47

Table 5e. Percent of Objective Writing Items Falling into Each Category

Test	Scoring Criteria Factors				
	Mechanics	Word Choice	Organization	Style	Insight
AP	X	X	X	X	X
CSU	X	X	X	X	X
GSE Reading/Literature			X		X
GSE Written Composition	X	X	X	X	X
SAT II Writing	X	X	X	X	
UC Subject A	X	X	X	X	X

Table 5f. Factors Identified in the Scoring Criteria of Each Test

a nonsense word was used in a sentence, and students were asked to decipher the meaning of the nonsense word. Unlike the other vocabulary tasks described earlier, students must infer the meaning of the word based on how it is used, and cannot rely on prior knowledge to answer the item.

Similarly, the SAT I contained questions assessing analytical and inference ability. The SAT I included an analogy section that required students to analyze the relationships between a pair of words, and choose another pair of words whose relationship was most similar to the original pair. The SAT I also contained an additional section in which a sentence with omitted words was presented. Examinees were then asked to choose which set of words, when inserted into the sentence, would make the sentence most meaningful. A unique feature of some of the SAT I items was that they addressed not only the primary meaning of a word, but the secondary and tertiary meanings as well.

Implications of the Misalignments

As with math, the misalignments among the ELA assessments can send confusing messages. There appeared to be little consistency among the exams, thereby rendering it difficult to counsel students on the best preparation methods. Measures that include only multiple-choice items would be approached in vastly different ways than exams that require a sample of the examinees' writing proficiency. Moreover, even when two tests require a written composition, the variations in the administrative conditions and scoring criteria call for different kinds of strategies. For instance, teachers sometimes instruct students to organize their thoughts

with a detailed outline. This technique may be appropriate for a two-hour UC essay, but it is less feasible for a 20-minute SAT II Writing task. Again, it is important to acknowledge that some of these inconsistencies may be more problematic than others, given the diverse purposes and examinees populations of these testing programs.

The inconsistency in the scoring rubrics, particularly the omission of mechanics, word choice, and style from the GSE Reading/Literature scoring rubrics and of insight from the SAT II Writing scoring guidelines, give rise to several concerns. First, these skills are part of the scoring criteria in most English courses and for the other assessments we examined. This means that the GSE Reading/Literature and SAT II Writing standards are incongruent with those that are typically expressed. Additionally, it is highly unlikely that the raters would be unconcerned with these factors when scoring the test, as mechanics, word choice, style, and insight are inherently part of what constitutes good writing ability. If raters are indeed including these skills as part of the scoring criteria, then students have been misinformed about the standards on which they are judged. In light of the kinds of signals the scoring rubrics send, developers of the GSE Reading/Literature and SAT II Writing assessments may wish to reconsider the current guidelines, and be more explicit about their scoring criteria.

Finally, there are concerns about the inconsistencies among the scoring standards across different measures of writing ability. The requirements for a model essay under the GSE Written Composition or CSU guidelines are less rigorous than those found for the AP exam.

For the two former tests, maximum scores were awarded to sample essays that had diction errors, usage and mechanics lapses, and underdeveloped paragraphs. Under the AP guidelines, such compositions might receive adequate scores, but would not be viewed as exemplary; only essays that demonstrate exceptional rhetorical and stylistic techniques, with substantial evidence to support a position, would receive a maximum score under the AP scoring rubrics. Because the GSE Written Composition, CSU, and AP exams are intended for different student ability levels and serve different purposes, misalignments among their scoring criteria are inevitable. Nevertheless, such discrepancies may send mixed messages to students and school personnel regarding the standards of what is considered an excellent composition.

Discussion

In general, many of the studied tests were not well-aligned with respect to structure or content. However, whether the inconsistencies are a source for concern needs to be interpreted in light of the purpose to which the assessments are intended. The misalignments may not pose a problem if they represent legitimate differences stemming from diverse uses of the measures. Indeed, different test purposes will necessitate different kinds of formats, administrative conditions, and item content. As was discussed earlier, variations in the content and difficulty level of the SAT Level IIC, Stanford 9, and GSE math tests should not be considered problematic, as the exams have different test uses, and it is virtually impossible to create one test that can simultaneously serve those different purposes. However, when the measures serve

similar purposes and examinee populations, yet differ substantially in terms of content and cognitive demands (as appears to be the case for the GSE High School Math and SAT Level IC assessments, for example), there may be valid concerns regarding the misalignments.

Regardless of whether or not the discrepancies are warranted, the inconsistencies can translate to a perceived testing overload by the examinees. Consider, for instance, the students applying for entrance to the University of California system. They are required to take the SAT I or ACT, SAT II, and possibly a placement exam. They are also encouraged to take the GSE and AP exams. The overabundance of exams students are required to take can foster a perception that the various measures are redundant. Although many of the tests have distinct uses and are therefore not interchangeable, it is likely that many students will not recognize the reasons underlying the need for multiple assessments, and may view the exams as unnecessary, time-consuming, and stressful.

The misalignments can also send inconsistent signals with respect to preparation efforts. Although all of the testing preparation materials claimed that a challenging and rigorous academic program was the best way to prepare for their exams, structural and content variations among the tests dictated differences in the most appropriate preparation strategies. It is likely that instructors confronted with preparing students for the entry-level CSU placement exam would most likely approach this task in a different manner than if they were to prepare their students for the more rigorous college-entrance assessments. Perhaps the most important signaling function of the tests relates to the messages they send to students about what kinds of

skills are valued. It has been shown that large-scale assessments, particularly those with direct consequences for students or teachers, often influence the kinds of skills and knowledge that are developed.¹⁴ That is, both students and teachers are likely to focus their attention on the content that is tested. For this reason, there have been efforts from various educational reform movements and professional development organizations to increase the emphasis given to problem-solving items that are framed in real-world contexts.

However, there remains a disassociation between the skills that are considered valuable and the skills that are actually assessed. In math, the majority of the items on the studied assessments involved heuristics using procedural or declarative knowledge. Moreover, as few items had meaningful applications to the real world, these tests do not convey the importance of math beyond the classroom or testing context. It appears that despite efforts to the contrary, students may be receiving messages that mathematics is a sequence of algorithms to be memorized and applied, with little connection to real life problems.

Similarly, on the ELA assessments students are not given clear signals as to which skills are valued. Arguably, the ability to make inferences or to evaluate the style of a given piece is as valuable as the ability to remember information, but this message is probably not adequately conveyed by exams such as the Stanford 9 or the GSE Reading/Literature test. Such tests encourage students to direct their efforts toward recollection of facts and details, as opposed to deeper analysis of the given passage. Moreover, the emphasis given to recall skills, particularly with respect to the assessment of

vocabulary, can lead some students to learn the definitions of words through rote methods, such as memorization. Although this may lead to an increase in scores, it is not the ideal way of acquiring meaning, as nuances are not learned as adequately as if the word had been encountered in context.

Perhaps the most problematic signal arises from the exams administered at the high school level, including the Stanford 9 and the exams required for college admissions, because the majority of these do not require examinees to demonstrate their writing skills. The SAT II Writing test, which does include an essay item, does not require multiple writing samples, nor does it allow an extended period of time for students to develop their ideas fully in a single essay. This may serve to communicate to students that writing is not an essential skill for college-level courses. In reality, however, most university-level classes require students to write extensively. Thus, the kinds of skills and knowledge valued in universities can differ substantially from students' expectations. Again, if the measures are to send signals that writing ability is a desired skill, then the current tests need to be modified to reflect that message.

Limitations of the Alignment Analysis and Recommendations for Future Work

The use of expert judgments is a fairly common approach to studying alignment as well as content validity.¹⁵ The evidence gathered through this study will be useful in evaluating the validity of currently used tests for the purposes for which they were designed. However, this study

does not provide a complete picture of these assessments, and other analytic approaches might lead to somewhat different conclusions. Observations and interviews with students as they take the tests, an approach that is sometimes used during the test development process, would undoubtedly result in somewhat different interpretations of a tests reasoning requirements. Empirical data are also needed to quantify the consistency of student performance across various kinds of tests. It is important to evaluate the likelihood that students who perform well on one kind of assessment will do so on another, as large discrepancies in performance can send confusing signals regarding the actual proficiency level of a student. Particularly for examinees attempting to prepare for a high-stakes measure, it is essential that they receive accurate and consistent information about their strengths and weaknesses. Finally, increasing the number of forms studied for each assessment would enhance the generality of our findings. The studied tests represent a sample of skills from a single testing occasion, and forms from other occasions will certainly vary somewhat. This is especially true when we analyze alignment among ELA topics, where there is a limited sample on any given test form (e.g., there may be only one essay). Studying multiple forms could increase the stability of our results.

The study would also be improved if we increased the number of expert raters and refined our analysis of agreement levels among these raters. An ideal study would bring in a larger number of expert judges, selected to represent a range of experience in both the K-12 and higher education sectors. It would also involve a more systematic analysis of differences in coding, with perhaps some quantifica-

tion of commonalties and differences among tests. Because we looked at a large number of tests across several states,^{iv} it was not feasible to conduct a more thorough study. However, as we argued earlier, alignment is a more critical consideration for some sets of tests than for others. Therefore there may be great benefit in conducting a more comprehensive alignment study on the few tests for which alignment really matters, allowing resources to be targeted rather than spread across a large number of tests. In California, for example, it would be worth conducting a study in which the Stanford 9 test is compared with other measures of high school math and reading achievement, such as the SATII exams. Comparisons with the SATI are arguably less relevant. In any case, it is clear that students, parents, educators, and policymakers all could benefit from attention to the messages and signals that tests are sending students.

An additional problem stems from the lack of availability of full test forms for some of the testing programs. Inspection of actual forms would provide more accurate information about the distributions of items across our various categories. On the other hand, because this study is focused on the signaling function of tests, the use of publicly released materials rather than actual forms may actually be preferable. It is unlikely that students remember many details of the items they took on a single testing day. In contrast, the preparation materials, including sample items and sample test forms, probably have a greater influence on students preparation behaviors and their interpretation of what the test measures.

Finally, many of the interpretations we make above depend on assumptions about stu-

dents interpretations of the signals sent by tests. It would be extremely valuable to interview students, educators, and other school and college personnel to assess their views on these various testing programs and to find out how the tests influence their teaching and learning. It is also important to discover whether some groups of students are more

heavily influenced by these tests than are others. For example, the group of students who engage in extensive SAT preparation activities is undoubtedly different from those who take the SAT with little prior preparation. Data collected as part of the Stanford Bridge Project will provide useful information to supplement this alignment study.

Notes

i We did not include the results for the AP Calculus AB exam because it was markedly different from the other studied tests. For example, it did not include material from any other mathematical content area except calculus, and was the only measure that necessitated a graphing calculator. Moreover, it was intended to assess the proficiency level of a very select group of high-ability students. Given that the AP shared few commonalities with the other assessments, it was excluded from the following discussion.

ii Again, for the GSE we did not examine an actual test form, but instead use the set of released items given to teachers and students. Thus the percentages discussed here do not represent percentages of items that examinees take, but instead indicate the relative emphases given to various topics on the materials that students use to prepare for the tests.

iii As discussed earlier, whether an item assesses inferential skills or recall ability depends upon a student's proficiency level.

iv Although this report includes only California, we performed similar analyses for five additional states.

1. *Making Standards Matter 1997: An Annual Fifty-State Report on Efforts to Raise Academic Standards*. (Washington, DC.: American Federation of Teachers, 1997).

2. See, for example, M.W. Kirst, *Improving and Aligning K-16 Standards, Admissions, and Freshman Placement Policies*. (Stanford, CA: National Center for Postsecondary Improvement, 1998); A. Venezia, "Connecting California's K-12 and Higher Education Systems: Challenges and Opportunities," this volume.

3. R. Edgerton, *Higher Education White Paper*. (The Pew Charitable Trusts: 1997).

4. M. Smith, M., and J. ODay, "Systemic School Reform, in *The Politics of Curriculum and Testing*," eds. S. H. Fuhrman & B. Malen (Bristol, PA: The Falmer Press, 1991), 233-268.

5. *National Education Summit Briefing Book* (Achieve, Inc., 1999), available at <http://www.achieve.org>.

6. N.L. Webb, *Alignment of Science and Mathematics Standards and Assessments in Four States*, Research Monograph No. 18 (Madison: National Institute for Science Education, 1999).

7. *Not Good Enough: A Content Analysis of Teacher Licensing Examinations* (Education Trust, Spring 1999).

8. L. Burstein, D. Koretz, R. Linn, B. Sugrue, J. Novak, E. Baker, and E.L. Harris, Describing Performance Standards: Validity of the 1992 National Assessment of Educational Progress Achievement Level Descriptors as Characterizations of Mathematics Performance, *Educational Assessment* 3 (1995/1996): 9-51; Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the Trial State Assessment: An Evaluation of the 1992 Achievement Levels (Stanford, CA: The National Academy of Education, 1993).

9. B.M. Stecher, and S.I. Barron, Quadrennial Milepost Accountability Testing in Kentucky, CSE Report 505 (Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, 1999).
10. Kenny and Silver, 1999.
11. Ibid.
12. National Assessment Governing Board, *Mathematics Framework for the 1996 National Assessment of Educational Progress: NAEP Mathematics Consensus Project*, GPO 065-000-00772-0 (U.S. Department of Education, 1995), available at <http://www.nagb.org>.
13. *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: National Council of Teachers of Mathematics, 1989).
14. G.F. Madaus, "The Influence of Testing on the Curriculum, in *Critical Issues in Curriculum*," ed. L. Tanner (Chicago: University of Chicago Press 1988), 83-121; T. Kellaghan, and Madaus, "National Testing: Lessons for America from Europe," *Educational Leadership* 49 (1991): 87-93; Stecher and Barron, 1999.
15. S.G. Sireci, "Gathering and Analyzing Content Validity Data," *Educational Assessment*, 5 (1998): 299-321.