

This document contains *Chapter 2: Review of the Test* from the Third Biennial Report, California High School Exit Examination (CAHSEE) published on February 1, 2006, by the California Department of Education. The entire report is available at <http://www.cde.ca.gov/ta/tg/hs/thirdbiennial.asp>.

## Chapter 2: Review of the Test

### *Introduction*

As part of the ongoing evaluation of the California High School Exit Exam (CAHSEE), HumRRO conducted item review workshops in June 2005 with California content experts in mathematics and English-language arts. This chapter presents the results of the two workshops, one held in northern California and one held in southern California.

### *Earlier Independent Item Reviews*

The 2005 item review workshops involved two related activities to monitor the quality and accessibility of the CAHSEE. In particular, HumRRO carried out investigations of: (a) the degree of alignment between the CAHSEE test questions (items) and the academic content standards, and (b) the degree of accessibility of the test questions and format for various student populations by examining elements of universal test design. An alignment study evaluates the extent of content match between the test questions and the content standards, examining whether the material on which students are assessed is the same as what is specified in the content standards. A universal test design study examines a test for appropriate format, scope, and content relative to the range of students who will be taking that assessment, such as students with limited English language proficiency and students with disabilities. The results of both types of investigations contribute to the assessment of the validity of the test as a measure of the targeted content.

The 2005 workshops extended results from CAHSEE item review workshops that we conducted in 2000 and 2002. The purpose of the 2000 workshop was to examine the alignment of the newly developed field-test items against the content standards and classroom curriculum. In that workshop, educators from California assessed the alignment of the test questions to their intended content standards by rating the degree of match between them. Overall, these reviewers determined that approximately 77 percent of English-language arts (ELA) items and 92 percent of math items matched well with the content standards for which they were developed. At that time, test blueprints had not yet been approved and test forms had not yet been constructed. We thus concluded that the test item pool as a whole represented the standards well. Reviewers also evaluated whether students would be able to answer the items based on their school's curriculum. In this case, the panelists found that the majority of items (90% for ELA and 65% for math) might be problematic for students based on the curriculum they received at that time. As a result, HumRRO recommended that curriculum specialists focus on bringing the curriculum more in line with the targeted content standards. More complete information is provided in Wise et al. (June 2000).

In 2002, the workshop panelists focused on the alignment of more recent CAHSEE test items with the content standards, and they compared the quality of these items with items in the 2000 review, many of which had become operational (i.e., used

## **Independent Evaluation of the CAHSEE: Third Biennial Report**

---

in calculating scores). Panelists used a rating system similar to the one used to evaluate alignment for the 2000 workshop. They determined that approximately 81 percent of ELA items and 83 percent of math items matched well with their target content standards. Thus, reviewers judged the ELA items to align slightly better than in the 2000 review, while they judged the math items to align less well than in the previous survey, but still better than the ELA items. Alignment for ELA likely improved due to more specific scoring rubrics developed for the essay items. There was some variation in the alignment ratings across the specific content areas within each subject. For ELA, the lowest alignment ratings were for items measuring literary response and analysis (71% strong alignment). For math, lower alignment ratings were found for mathematical reasoning (50% strong alignment) and the seventh grade statistics, data analysis, and probability items (67% strong alignment). Panelists were asked reasons for low alignment ratings of specific items. One common response for the ELA items is that they measured skills that were foundational for the intended target, but at a lower depth of knowledge. See Wise et al., (June 2002) for more complete information.

### ***Goals of the 2005 Item Review***

Our 2005 item review provided an opportunity to address questions that arose with the revision to CAHSEE test specifications introduced in 2003–2004, when the exam was restarted for the Class of 2006. The Board made slight adjustments to the test blueprints and the test developer was released from the requirement of matching closely the difficulty of each new test form to the difficulty of the original 2001 test form. The result of these revisions was a somewhat easier math test and a slightly more difficult ELA test. The Board also reset the performance level standards by keeping them at the same percent correct level (55% for math and 60% for ELA) as before. The result was that more students passed the math test than would have with the previous versions (Wise et al., 2004). Questions arose as to whether the revised math test was better, with items focused more closely on specific requirements, or was weaker, because the questions did not assess the full depth of the math standards.

Another key question concerned whether the questions provided a fair assessment for English learners and students with disabilities. Passing rates for these groups have been consistently lower than for other students. It was important to determine whether part of the performance gap might have resulted from features of the test questions that made them inappropriately difficult for these students.

In the 2005 item review workshops, HumRRO adopted recently developed methods to assess both alignment and item quality. For the alignment process, we used the method created by Norman Webb (1997; 1999; 2005) and the Council of Chief State School Officers (CCSSO). In addition, we asked the National Center on Educational Outcomes (NCEO) to provide their expertise on universal test design in the review of test accessibility (see *Considerations for Universally Designed Assessments*, NCEO, 2005).

Both of these activities also provided evidence of meeting requirements of the No Child Left Behind Act (NCLB) of 2001. In the document on *Standards and Assessment Peer Review Guidance* (April, 2004), the U.S. Department of Education requested evidence of the alignment and accessibility of each state's assessment systems. In particular, this document stipulated that:

- “Assessments must be aligned with State academic content and achievement standards, and they must provide coherent information about student attainment of State standards in at least mathematics and reading/language arts.
- The same assessment system must be used to measure the achievement of all students.
- The assessment system must be designed to be valid and accessible for use by the widest possible range of students, including students with disabilities and students with limited English proficiency (LEP).” (pp. 2–3)
- The original NCLB legislation also points to the need for an inclusive test design. Specifically, it requires that all assessments “be designed from the beginning to be accessible and valid with respect to the widest possible range of students, including students with disabilities and students with limited English proficiency” (NCLB, Section 200.2(b)(2)).

The alignment and universal test design results are discussed in detail in two separate sections of this chapter. While both of these activities occurred within the same workshop, the method and analyses for alignment and universal test design involve distinct processes. The first part of this chapter discusses the alignment methods, results, and subsequent recommendations in the section entitled “*Item Review Workshops: Alignment of the CAHSEE to the Academic Content Standards*”. The second part of the chapter presents the methods, results, and recommendations for universal test design in the section “*Item Review Workshops: Universal Test Design of the CAHSEE*.”

### ***Item Review Workshops: Alignment of the CAHSEE to the Academic Content Standards***

For the alignment tasks, HumRRO evaluated the level of content agreement between the CAHSEE test questions and the targeted mathematics and English-language arts standards. As a preface to the discussion of the alignment tasks and results, we first describe several core concepts related to assessment and alignment research.

#### ***Assessment-to-Standards Alignment***

The term *alignment* refers to “the degree to which [content] expectations and assessments are in agreement” (Webb, 2005). Alignment analyses indicate the breadth, or scope, of knowledge included in the assessment. In addition, alignment analyses examine the depth of knowledge, or cognitive processing, required of students by the assessment compared with the state's content standards. In other words, alignment

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

analyses help to answer questions such as, “How much content is covered by the assessment?” “Is this content sufficiently similar to the expectations of the standards?” and “Are students asked to demonstrate this knowledge at the same level of rigor as expected in the content standards?”

Alignment concerns should be addressed early in the item development process. In fact, ETS has implemented a number of processes, from item writer training and guides through numerous reviews, to ensure that all items measure targeted content appropriately. The study reported here was not an attempt to review specific item development and review procedures employed by the test developer. Rather, it was an independent check of the test questions that come out of the end of these processes.

### **California Academic Content Standards and Test Blueprints**

The CAHSEE test blueprints list a subset of the *California Academic Content Standards* identified by the High School Exit Exam (HSEE) Panel and approved by the Board as critical knowledge and skill for high school graduation. The CAHSEE blueprints draw on the full set of California Academic Content Standards across a range of grades and assign target numbers of test items to be included for each selected content standard.

For purposes of the alignment workshop, content standards were combined across grades and organized into the major content groupings shown in Table 2.1.

**Table 2.1. Content Strands Assessed by the CAHSEE for ELA and Math**

English-Language Arts Strands	Mathematics Strands
1. Word Analysis	1. Statistics, Data Analysis, and Probability
2. Reading Comprehension	2. Number Sense
3. Literary Response and Analysis	3. Algebra and Functions
4. Writing Strategies	4. Measurement and Geometry
5. Writing Applications (Genres and Their Characteristics)	5. Mathematical Reasoning
6. Writing Conventions	6. Algebra I

### ***Webb Alignment Method***

Several methods of alignment are in current use. Most methods involve ratings of several aspects of the assessment items relative to the content standards. The ratings are analyzed statistically to determine the extent of alignment. HumRRO used the alignment method developed by Norman Webb (1997; 1999; 2005) to evaluate the CAHSEE. The *Webb Alignment Method* includes specific criteria for judging the quality of alignment. Recent work to extend this method was supported by the Council of Chief State Schools Officers (CCSSO) and, as a result, Webb's method has been used widely in other states. We present below some explanation of terminology related to Webb's method before describing the specifics of this method.

***Standards Levels.*** The terminology used to describe specific expectations for student achievement varies widely from state to state. For this workshop, we adopted Webb's terminology describing different levels of content organization. Based on evaluations of a number of states, Webb has found that standards documents generally are divided into two or three organizational levels. Webb labeled these common levels as: (a) standard, (b) goal, and (c) objective. A *standard* is the highest, most general level of the content expectations, often written as a broad content category. The results of the analyses are reported at the standard level, meaning how well the test items align with each of these broad content categories (Webb, 2005). Table 2.1 above lists the broad content categories used with the CAHSEE content standards.

Standards documents always include at least one additional level with more specific content expectations. A *goal* is the next (middle) level of the content expectations. The goal includes smaller topics or subcategories within the standard, often written as general performance expectations. Not all standards documents include a goal level. The CAHSEE standards for both mathematics and English-language arts do not always delineate a content expectation at this level; this absence does not affect the outcomes on the degree of alignment.

Finally, the *objective* is the lowest, most specific level of the content expectations. These statements identify individual tasks and knowledge expectations at a more detailed level than the goal or standard levels. Since assessment items are written at this level of specificity, reviewers rate items at the level of the objective per standard.

The *California Academic Content Standards* for English-language arts are organized around four levels including domain, strand, substrand, and standard (from most general to most specific). For consistency, Webb's labels and meaning for content expectations were applied to the California Academic Content Standards as shown in Table 2.2.

**Independent Evaluation of the CAHSEE: Third Biennial Report**

---

**Table 2.2. Webb Labels Applied to California Academic Content Standards for English-Language Arts**

<b>Current Labels for California Academic Content Standards for English-Language Arts:</b>			
<b>Domain</b>	<b>Strand</b>	<b>Substrand</b>	<b>Standard</b>
Reading	1.0 Word analysis, fluency, and systematic vocabulary development	Vocabulary and concept development	1.1 Identify and use the literal and figurative meanings of words and word derivations.
<b>Webb Labels Applied to California Academic Content Standards for English-Language Arts:</b>			
<b>Subject</b>	<b>Standard</b>	<b>Goal</b>	<b>Objective</b>
Reading	1. Word Analysis	Vocabulary and concept development	1.1 Identify and use the literal and figurative meanings of words and word derivations.

The California content standards for mathematics generally are organized into strands and standards. A *strand* refers to a broad content category, while *standard* refers to specific statements of content expectations. In contrast to Webb, the term *standard* for California refers to the most specific level of the content expectations. Again, these content expectations were relabeled to match Webb’s method more closely as shown in Table 2.3.

**Table 2.3. Webb Labels Applied to California Academic Content Standards for Mathematics**

Current Labels for California Academic Content Standards for Mathematics		
Strand	Standard	
Grade 7-Number Sense	1.0 Students know the properties of, and compute with, rational numbers expressed in a variety of forms:	1.1 Read, write, and compare rational numbers in scientific notation (positive and negative powers of 10) with approximate numbers using scientific notation.
Webb Labels Applied to California Academic Content Standards for Mathematics		
Standard	Goal	Objective
2. Number Sense	1.0 Students know the properties of, and compute with, rational numbers expressed in a variety of forms:	1.1 Read, write, and compare rational numbers in scientific notation (positive and negative powers of 10) with approximate numbers using scientific notation.

**Webb Alignment Criteria.** The Webb method evaluates alignment between assessments and standards by measuring four criteria:

1. Categorical concurrence
2. Depth of knowledge consistency
3. Range of knowledge correspondence
4. Balance of representation

For a complete analysis of alignment, all four of Webb’s criteria must be considered together. However, each criterion provides different information about the degree of alignment between the assessment and content standards. A brief description of each criterion is presented here. We provide more detailed information on the statistical indicators used with each of these criteria in our 2005 Evaluation Report (Wise, et al., 2005) and also by Webb (2005).

**Categorical concurrence** is a basic measure of alignment between standards and test items. This term refers to the proportion of overlap between the content stated in the standards and that assessed by items on the test. It is assessed by counting the number of items judged by experts to be good matches to each targeted standard. Webb maintains that standards should be assessed by a minimum of six items for acceptable categorical concurrence.

**Depth of knowledge (DOK)** measures the type of cognitive processing required by items and standards. For example, is a student expected to simply identify or recall basic facts, or is the student expected to use reasoning by manipulating information or strategizing? The purpose of using depth of knowledge as a measure of alignment is to determine whether the item and corresponding standard are both written at the same

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

level of cognitive complexity. Reviewers make separate judgments about cognitive complexity of the objectives and of the test items. These two judgments are compared to determine whether the items are written at the same level as the standard to which they are linked. Webb refers to his comparison as *depth of knowledge consistency*.

Another measure examines the ***range of knowledge correspondence*** between the test items and content standards. The range of knowledge measure looks at the breadth of coverage of the specific objectives under each standard. Webb (1999) requires that only a single item be linked to an objective in order for that objective to be counted as covered. Webb suggests that at least 50 percent of the objectives for a standard should be matched with one or more items to demonstrate acceptable range-of-knowledge correspondence.

Finally, the ***balance of representation*** criterion focuses on content coverage in yet more detail. Webb (1999) suggests that items should be distributed in an even way across the objectives for a standard to have good balance.

### ***Alignment Workshop Methods and Procedures***

To obtain a geographically representative sample of California educators, HumRRO conducted two separate workshops, one in northern California and one in southern California. The first day of each workshop was devoted to alignment evaluation, while the second day was devoted to universal test design. HumRRO staff conducted both workshops in the same way, using identical procedures and materials (e.g., rating forms).

***Workshop Participants.*** We contacted a total of 310 districts to recruit content experts for participation in the workshops. In addition, we made direct contact with 80 school administrators and 30 teachers. A considerable effort was made to represent experience with various student groups (e.g., English learners, students with disabilities). These contacts yielded a total of 26 teachers and curriculum specialists who participated in the item review workshops.

Of these panelists, one individual in each workshop served as a point of reference regarding students with specific physical impairments. In the northern workshop, a representative from a California School for the Blind fielded questions concerning the abilities and expectations for visually impaired students. In the southern workshop, a representative from a California School for the Deaf fielded questions related to hearing impairments. These two individuals did not serve as reviewers in the alignment analysis so that they could be available for both the math and ELA content groups.

Table 2.4 lists the number of remaining panelists who served as alignment reviewers by content area and current position.

**Table 2.4. Panelists by Content Area and Current Position**

Current position	English-language arts panelists	Mathematics panelists
Teacher, regular classroom	5	4
Teacher, special education	3	2
Teacher, EL	1	0
Curriculum Specialist	3	7
Total:	11	13

Table 2.5 includes the years of experience for these panelists. This information is broken down by region.

**Table 2.5. Experience Level of Panelists**

Content Area	Region	Less than 5 years	5–9 years	10–19 years	20 or more years
ELA	Northern	1	3	1	2
	Southern	0	1	1	2
Math	Northern	0	1	2	1
	Southern	0	1	5	3
Total		1	6	9	8

**Materials.** Reviewers evaluated the alignment between the assessments (mathematics or English-language arts) and their corresponding standards using Webb’s alignment methods and rating forms.

**Test Forms.** Reviewers assessed the February 2005 test form of the CAHSEE for English-language arts and mathematics. The test developer, ETS, provided HumRRO with a copy of these test forms as well as the item specifications. Table 2.6 presents the general format for each test.

**Table 2.6. Test Item Composition by Content Area**

Content Area	Total Items	Core Items	Field Test Items	Selected Response Items	Constructed Response Items
ELA	80	73	7	79	1
Math	92	80	12	92	0

Similar to most standardized assessments, the February 2005 test form includes both core items and field test items. Field test items include those items that are being evaluated for use on future exams, while core items are used to score the students. The core items have been field tested previously. Since only core items are used to compute scores, alignment analyses focused on core items.

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

Blueprints. Reviewers compared the mathematics and English-language arts items from the February 2005 test forms with the CAHSEE test blueprints<sup>3</sup>. As explained earlier, the assessment was compared with the test blueprint to ensure a more fair evaluation of alignment.

The CAHSEE test blueprints for mathematics and for English-language arts include a set number of assessed standards, goals, and objectives (Webb's terminology). The total numbers of each are presented in Table 2.7. One particular standard for ELA, Writing Applications, varies per test administration in the specific objective(s) assessed.

**Table 2.7. Number of Standards, Goals, and Objectives for Math and ELA**

Content Area	Standards	Goals	Objectives
English-language arts	6	17	33
Mathematics	7	26	53

Rating Forms and Instructions. Reviewers used two rating forms to make judgments about the standards and the assessment items separately. For the CAHSEE blueprints, reviewers used the Depth-of-knowledge (DOK) Rating Sheet to evaluate each assessed content objective. For the assessment items, reviewers used the Item Rating Sheet to evaluate each item on DOK and the primary and secondary content objectives linked with the item. See Wise et al., (2005) for more detailed information on the rating forms and instructions.

To perform the alignment task, reviewers received a copy of the Alignment Instructions and Definitions sheet. This sheet explained how to use each rating form with several examples. The sheet also included definitions for each DOK level, as shown in Table 2.8.

**Table 2.8. Depth of Knowledge Levels from Alignment Instructions Sheet**

Level	Title	Description
Level 1	Recall	Item requires simple recall of information, such as facts, definition, terms, or procedures.
Level 2	Skills/Concepts	Item calls for engagement in some mental processing and decisions beyond habitual response.
Level 3	Strategic Thinking	Item requires students to reason, plan, and use evidence.
Level 4	Extended Thinking	Item requires complex reasoning, planning, and thinking, typically over an extended period of time.

---

<sup>3</sup> The CAHSEE test blueprints for mathematics and English-language arts can be found on the CDE Web site. These blueprints were approved by the State Board of Education July 9, 2003.

***Debriefing Form.*** Reviewers completed Webb's debriefing survey at the end of the alignment tasks. This survey requested reviewers' overall impressions of the degree of alignment in a series of five questions.

### ***Alignment Results***

We begin with an analysis of the extent to which workshop participants agreed with the test developers as to the standards and objectives assessed by each test question. Next, alignment results are reported for each of Webb's four criteria. Again, we emphasize that Webb's terminology is used due to the structure of his analyses. Specifically, we refer to *standard*, *goal*, and *objective* in substitution of the California terms *strand*, *substrand*, and *standard*. However, the hierarchy (from broadest to most specific content expectation) is the same.

At the end of the Alignment Results section, we also include a brief summary of reviewers' comments. Reviewers were given the opportunity to make notations about items during the item rating period. In addition, they completed the Debriefing Survey, which asked for impressions about overall alignment.

***Rater Agreement Levels.*** Each test question is targeted to a particular standard and objective by the test developer. The objective-level assignments are used in test development to ensure that each form follows the test blueprint in terms of the number of items measuring each objective. The assignment of items to test standards (strands) is particularly critical as they determine which items are used in reporting information at the subscale level.

The Webb alignment process does not include assessing the extent to which reviewers' placement of items agrees with the operational placement of the test items. Before turning to the results of the Webb process, we provide a brief analysis of the agreement of the workshop participants with the operational placement of each item. Table 2.9 shows the percent of time the standard and objective matched by our raters agreed with the assignment of the test developer.

The raters generally agreed with the placement of the items with respect to the standards used in subscale reporting, but frequently disagreed with the particular objective within that standard that the item assesses. The lowest agreement rates were for the essay question, treated here as a single item under writing applications. Most of the reviewers believed that the essay also measured objectives under Writing Strategies. Also, only one essay is included in each form and so not all objectives under Writing Applications are covered. Reviewers consistently wanted to assign the essay to additional objectives and the result was a very low agreement rate at the level of objectives. Reviewers also linked some of the Writing Strategies items to objectives under Written and Oral English Language Conventions.

For mathematics, the agreement rates were generally higher. The primary area of disagreement was under Algebra and Functions, where some reviewers linked items to objectives targeted operationally for Algebra I objectives.

**Independent Evaluation of the CAHSEE: Third Biennial Report**

**Table 2.9. Agreement of Workshop Participants with the Operational Standards and Objectives Assigned by the Test Developer**

Standard Number	Standard (Strand)	Targeted Number of Items	Percent of Raters Assigning the Targeted Standard	Percent of Raters Assigning the Targeted Objective
ELA				
1	Reading-Word Analysis, Fluency, and Systematic Vocabulary Development	7	79%	79%
2	Reading Comprehension (Focus on Informational Materials)	18	66%	35%
3	Reading-Literary Response and Analysis	20	85%	53%
4	Writing Strategies	12	67%	26%
5	Writing Applications (Genres and Their Characteristics)	1	52%	52%
6	Written and Oral English Language Conventions	15	87%	87%
Overall Total		73	76%	53%
Mathematics				
1	Statistics, Data Analysis, and Probability	13	87%	76%
2	Number Sense	17	86%	69%
3	Algebra and Functions	20	73%	57%
4	Measurement and Geometry	18	91%	67%
6	Algebra I	13	85%	67%
Overall Total		80	87%	76%

Note: Mathematics reasoning items were also targeted to one of the above five content areas. These items are included under their primary content designation in the table above to avoid duplication. This increases the item counts for some strands above the minimum specified in the exam blueprints.

**Categorical Concurrence.** Categorical concurrence is a basic measure of alignment between standards and test items. This measure indicates how much general emphasis each standard receives on an assessment. Table 2.10 shows the results for ELA and for math averaged across reviewers from each workshop. The table lists the number and title of the standard, the target number of items listed in the test blueprint, the average number of items matched by reviewers, the standard deviation across reviewers in the number of items matched, and the conclusion of this alignment analysis. The bottom row under each content area indicates the total number of items included in the blueprint and matched by reviewers.

**Table 2.10. Categorical Concurrence: Average Number of Core Items per Standard**

Standard Number	Title of Standard	Number of Items Per Standard			At Least Six Items
		Target Number	Average Matched	Standard Deviation	
ELA					
1	Reading—Word Analysis, Fluency, and Systematic Vocabulary Development	7	8.36	2.62	YES
2	Reading Comprehension (Focus on Informational Materials)	18	10.55	3.36	YES
3	Reading—Literary Response and Analysis	20	20.09	5.11	YES
4	Writing Strategies	12	10.36	4.15	YES
5	Writing Applications (Genres and Their Characteristics)	1	1.00	1.34	NO*
6	Written and Oral English Language Conventions	15	14.18	5.21	YES
Overall Total		73	64.55	6.78	
Percent of standards with at least six items					83%
Mathematics					
1	Statistics, Data Analysis, and Probability	12	10.69	1.70	YES
2	Number Sense	14	14.69	2.18	YES
3	Algebra and Functions	17	16.15	3.02	YES
4	Measurement and Geometry	17	17.85	2.82	YES
5	Mathematical Reasoning	8	3.31	1.80	NO**
6	Algebra I	12	13.62	2.66	YES
Overall Total		80	76.31	2.42	
Percent of standards with at least six items					83%

\*Note. This standard corresponds with the writing item. The item links with several objectives within the standard as intended in the test blueprints.

\*\*Note. Mathematical reasoning is a process rather than a content area. Items that assess mathematical reasoning also assess one of the other content standards.

**English-Language Arts.** For ELA, Table 2.10 shows that the average across raters for the standard *Reading—Word Analysis, Fluency, and Systematic Vocabulary Development* is 8.36 items with a standard deviation of 2.62. This finding agrees closely with the blueprint target for this standard, which is 7 items. In comparison, the average number of items matched to the standard *Reading—Literary Response and Analysis* is 20.09 items with a standard deviation of 5.11. A higher standard deviation generally points to more variability in the ratings of each reviewer, which means that some reviewers' ratings are further away from the average. For example, the actual number of items matched to this standard by reviewers ranged from 4 to 25 items.

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

Based on these results, five of the ELA standards are represented adequately by the core items on the assessment. It should be noted that, while the standard *Writing Applications (Genres and Their Characteristics)* does not match a sufficient number of items based on the Webb method, this standard corresponds with the constructed response (essay) item. This outcome reflects the intended design of the test blueprint.

**Mathematics.** For math, the reviewers' item ratings met the minimum level of acceptable concurrence for five of six standards. For these five standards, the number of items matched the target numbers in the blueprints closely. The exception was *Mathematical Reasoning* (M = 3.31). For this content area, reviewers matched fewer items than were targeted in the blueprints.

*Mathematical Reasoning* is a complex standard to assess. All of the math items designed to assess reasoning ability also assess one of the content standards. Thus, there are number sense reasoning items, measurement and geometry reasoning items, and so on. As in prior reviews of CAHSEE items (Wise et al., 2000; Wise et al., 2002), the workshop participants were more likely to match these items to the content category rather than to this "process" standard. Difficulties in developing a clear specification of the reasoning process are not unique to this exam. *Further consideration should be given to the specification of objectives for this standard when revisions to the content frameworks are next considered.* Note that separate score information is not reported for mathematical reasoning, as it is for the other strands. Consequently, low categorical concurrence results for this standard are not as critical.

***Depth of Knowledge Consistency.*** *Depth of knowledge* (DOK) measures the type of cognitive processing required by items and content objectives. Table 2.11 includes the depth of knowledge consistency results for ELA and math. The table shows the percent of items judged to be below, at, or above the depth of knowledge of the corresponding content objective. The final column indicates whether the distribution of depth of knowledge ratings for items within each standard meet Webb's criteria that at least half of the items be at or above the level of the corresponding objectives.

**Table 2.11. Depth of Knowledge Consistency: Average Percent of Core Items with DOK Below, At, and Above DOK Level of the Corresponding Objective**

Standards	Average Items per Standard	Depth of Knowledge Consistency						DOK Consistency (min 50% of Items At or Above)	
		% Items Below		% Items At Same Level		% Items Above			
Title		M	S.D.	M	S.D.	M	S.D.		
<b>ELA</b>									
1	Reading—Word Analysis, Fluency, and Systematic Vocabulary Development	8.36	85	0.16	15	0.16	0	0	NO (15%)
2	Reading Comprehension (Focus on Informational Materials)	10.55	73	0.17	23	0.14	4	0.06	NO (27%)
3	Reading—Literary Response and Analysis	20.09	38	0.22	49	0.21	13	0.06	YES (62%)
4	Writing Strategies	10.36	55	0.30	39	0.25	6	0.11	NO (45%)
5	Writing Applications (Genres and Their Characteristics)	1.00	0	0	56	0.46	44	0.45	YES (100%)
6	Written and Oral English Language Conventions	14.18	48	0.26	38	0.22	14	0.29	YES (52%)
Overall Total		64.55	48	0.33	38	0.29	14	0.30	
Percent of standards with 50% of item DOK at or above objective DOK:									50%
<b>Mathematics</b>									
1	Statistics, Data Analysis, and Probability	10.69	39	0.19	51	0.17	10	0.16	YES (61%)
2	Number Sense	14.69	33	0.13	57	0.10	10	0.09	YES (67%)
3	Algebra and Functions	16.15	48	0.19	45	0.17	7	0.09	YES (52%)
4	Measurement and Geometry	17.85	37	0.20	51	0.16	12	0.09	YES (63%)
5	Mathematical Reasoning	3.31	33	0.28	61	0.33	6	0.15	YES (67%)
6	Algebra I	13.62	49	0.21	40	0.16	11	0.16	YES (51%)
Overall Total		76.31	35	0.14	52	0.07	14	0.10	
Percent of standards with 50% of item DOK at or above objective DOK:									100%

English-Language Arts. As shown in Table 2.11, the ELA reviewers found an acceptable level of consistency between the DOK levels of core items and corresponding objectives for three standards (numbered 3, 5, and 6 in Table 2.10). The DOK levels of items matched with the other three standards did not meet the minimum level of acceptable consistency, although the average depth of knowledge ratings for Writing Strategies items was close to the 50 percent minimum.

Mathematics. The average number of items at or above the DOK level of the objectives exceeded the 50 percent requirement for all six math standards.

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

***Range of Knowledge.*** Range of Knowledge measures how completely the test items cover the content objectives within each standard. The assessed objectives within a standard should be linked with at least one test question. Webb's minimum level of acceptability for range of correspondence is 50 percent per standard. This means that at least 50 percent of the objectives must be matched to an item.

Table 2.12 includes the results for ELA and math. This table includes the number of content objectives for each standard listed in the blueprints, the average number of items per standard (from Table 2.9), the average number of objectives linked with at least one item, and the conclusion for this alignment analysis. The bottom row lists the percent of standards with at least one item matched to 50 percent or more of the objectives.

***English-language arts.*** The ELA reviewers found that the core items linked with a sufficient number of objectives for five of the six standards. The standard *Writing Applications (Genres and Their Characteristics)* is supposed to be assessed by the single essay item, with coverage of different objectives rotated across forms. Thus, the essay from a single form did not meet the criteria for covering all of the Writing Application objectives. Across all of the ELA standards, 64% of the objectives were judged to be covered by at least one item, suggesting a sufficient range of knowledge for the test as a whole.

**Table 2.12. Range of Knowledge: Average Percent of Objectives per Standard Linked with Core Items**

Standards		Range of Objectives						Range of Knowledge Correspondence
Title	Number of Objectives	Average Items per Standard	Objectives with At Least One Item		% of Total Objectives per Standard			
			M	S.D.	M	S.D.		
			ELA					
1	Reading—Word Analysis, Fluency, and Systematic Vocabulary Development	2	8.36	1.91	0.30	95	0.15	YES
2	Reading Comprehension (Focus on Informational Materials)	6	10.55	4.00	1.10	67	0.18	YES
3	Reading—Literary Response and Analysis	12	20.09	7.55	1.57	63	0.13	YES
4	Writing Strategies	5	10.36	3.09	1.22	62	0.24	YES
5	Writing Applications (Genres and Their Characteristics)	5	1.00	1.50	1.14	30	0.04	NO
6	Written and Oral English Language Conventions	3	14.18	2.55	0.82	85	0.27	YES
Overall Total		33	64.55	21	3.09	64	0.08	
Percentage of Standards with 50% of Objectives Linked to At Least One Item								83%
Mathematics								
1	Statistics, Data Analysis, and Probability	7	10.69	4.69	0.85	67	0.15	YES
2	Number Sense	10	14.69	8.46	1.05	85	0.07	YES
3	Algebra and Functions	10	16.15	8.08	1.12	81	0.07	YES
4	Measurement and Geometry	10	17.85	8.31	1.11	83	0.09	YES
5	Mathematical Reasoning	6	3.31	1.83	1.03	31	0.18	NO
6	Algebra I	10	13.62	8.62	1.12	86	0.11	YES
Overall Total		53	76.31	40	3.15	75	0.23	
Percentage of Standards with 50% of Objectives Linked to At Least One Item								83%

**Mathematics.** The assessment was judged to adequately represent the range of content specified for five of the six mathematics standards. Approximately 40 of the 53 objectives across these standards were matched to core items.

The reviewers judged that the math items did not represent the range of knowledge well for the standard *Mathematical Reasoning*. As noted earlier, all of the items developed to assess mathematical reasoning were also designed to assess an objective under one of the content-specific standards. Reviewers tended to match these items to the content-specific objective and so coverage of mathematical reasoning appears sparse. Overall, reviewers judged 76 percent of the objectives to be covered by at least one test question, reflecting good range of knowledge coverage.

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

**Balance of Representation.** The fourth measure of alignment included in the Webb method is balance of representation. This criterion focuses on content coverage in greater detail. While the range of knowledge tells us something about the number of objectives that are linked to at least one test item, the balance measure takes into account how many items are linked with each objective per standard. Results showed adequate Balance of Representation for each of the standards assessed by the CAHSEE. (See Wise et al., 2005 for more detail on the assessment of balance of representation.)

**Reviewer Comments.** In addition to providing more standardized ratings of the core items, some reviewers gave written and verbal comments on the test items in space provided on their ratings sheets. These comments were passed on to the test developers for their consideration.

Reviewers also completed a *Debriefing Survey* to provide overall impressions on the degree of alignment. The survey, adapted from Webb (2005), includes four questions, as well as space for general comments. A summary of responses to the questions on mathematics is provided in Table 2.13. The comments represent individual responses for reviewers. Most responses and comments from these reviewers were positive, supporting the outcomes on the standardized ratings showing good alignment.

**Table 2.13. Debriefing Survey for Mathematics: Summary Responses**

Question	Response Options	Percent of Reviewers (N=13)	Comments/Opinions
1. For each standard, did the items cover the most important topics you expected? If not, what topics were not assessed that should have been?	YES	54% (7)	<ul style="list-style-type: none"> <li>• Concept of 'factoring' is the foundation for other concepts in algebra, but no items on this topic.</li> <li>• Seemed to be more items linked with algebra than listed in the blueprints.</li> <li>• Grade 6 Statistics was not covered.</li> <li>• A lot of emphasis on Number Sense.</li> <li>• Several basic algebra concepts were not covered.</li> </ul>
	MOSTLY	31% (4)	
	NO	15% (2)	
2. For each standard, did the items cover the most important performance levels you expected? If not, what performance was not assessed?	YES	23% (3)	<ul style="list-style-type: none"> <li>• Most items assessed at level 3 DOK.</li> <li>• Not all levels expected by a standard were covered.</li> <li>• Most items had lower DOK than expected in standards.</li> </ul>
	MOSTLY	69% (9)	
	NO	8% (1)	
3. Was there any content you expected to be assessed, but found no items assessing that content? What was that content?"	YES	23% (3)	<ul style="list-style-type: none"> <li>• No content assessed on functions.</li> <li>• More on algebra.</li> <li>• Some algebra was "light" on items.</li> <li>• Grade 7 Math Reasoning was assessed, but I had difficulty identifying which parts of the standard matched the items.</li> </ul>
	MOSTLY	8% (1)	
	NO	69% (9)	
4. What was your opinion of the alignment between the standards and assessment:			<ul style="list-style-type: none"> <li>• Most questions seemed to be written with a specific objective in mind.</li> </ul>
a. perfectly aligned.	a.	0%	
b. acceptable alignment.	b.	62% (8)	
c. needs slight improvement.	c.	38% (5)	
d. needs major improvement.	d.	0%	
e. not aligned in any way.	e.	0%	

Table 2.14 shows the responses and comments from the ELA reviewers. While alignment outcomes were acceptable overall, reviewers took issue with several specific features of the exam. In particular, a number of reviewers considered that the items assessed the elementary ELA standards more than the higher grade standards on which the CAHSEE is based. A second major theme pertained to exam accessibility for different types of students. Reviewers in the universal test design portion of the workshop reiterated these comments as well.

**Independent Evaluation of the CAHSEE: Third Biennial Report**

**Table 2.14. Debriefing Survey for ELA: Summary Responses**

Question	Response Options	Percent of Reviewers (N=11)	Comments
1. For each standard, did the items cover the most important topics you expected? If not, what topics were not assessed that should have been?	YES	18% (2)	<ul style="list-style-type: none"> <li>Some standards were “under-assessed”: 2.4, 2.5, 2.7, 3.1, 3.3, 3.5.</li> <li>Expected more questions asking students to synthesize reading.</li> <li>Items vaguely address standards because they are examples of elementary standards.</li> <li>I have a major concern that too many items are not aligned to our Grade 9–10 standards.</li> <li>Many items tested skills below grade level.</li> </ul>
	MOSTLY	45% (5)	
	NO	36% (4)	
2. For each standard, did the items cover the most important performance levels you expected? If not, what performance was not assessed?	YES	18% (2)	<ul style="list-style-type: none"> <li>Not a lot of items at DOK level 4.</li> <li>Items were hard to assess because a single objective could hit several different levels.</li> <li>Items seemed to ask students for lesser skills than the standards.</li> <li>Many questions assessed DOK 1 and 2.</li> </ul>
	MOSTLY	36% (4)	
	NO	45% (5)	
3. Was there any content you expected to be assessed, but found no items assessing that content? What was that content?”	YES	54% (6)	<ul style="list-style-type: none"> <li>Expected to see more on Reading 3.1</li> <li>No content that was missed, but there are flaws in the way the standards are written.</li> <li>Students were not often asked to “analyze”, “interpret”, or “explain”.</li> </ul>
	MOSTLY	0%	
	NO	46% (5)	
4. What was your opinion of the alignment between the standards and assessment:			<ul style="list-style-type: none"> <li>Several standards seemed to test elementary school standards—general content matched but not the specific objectives in this level of standards.</li> <li>There are too many reading passages, which take students a really long time.</li> <li>Items do not really align well with the higher order tasks of the Grade 9, 10 standards.</li> <li>No passages relate to experiences of minority, immigrant, urban students.</li> <li>Seem to be some cultural/disability biases.</li> <li>Concern for students with disabilities in taking this test.</li> <li>The exam aligns more with elementary standards rather than 8<sup>th</sup>, 9<sup>th</sup>, or 10<sup>th</sup>.</li> </ul>
a. perfectly aligned.	a.	0%	
b. acceptable alignment.	b.	9% (1)	
c. needs slight improvement.	c.	64% (7)	
d. needs major improvement.	d.	27% (3)	
e. not aligned in any way.	e.	0%	

**Alignment Conclusions**

The purpose of the 2005 alignment evaluation was to determine the level of content agreement between the February 2005 version of the CAHSEE and the designated California content standards for mathematics and English-language arts. Alignment between state academic standards and assessments is a requirement of the No Child Left Behind Act of 2001. This study serves as evidence of assessment-to-standards alignment for the CAHSEE.

Using the Webb method of alignment, HumRRO determined that the February 2005 test form did align with the content standards as specified in the test blueprints. As with many other states, the specific degree of alignment with the standards varied some per content area. Thus, California may wish to consider a review of those elements of the CAHSEE that aligned to the standards at lower levels. Such a review would be reasonable given the purpose of the CAHSEE as a high-school exit exam.

Table 2.15 provides a summary of the alignment outcomes for mathematics and for English-language arts. Based on Webb’s method, separate alignment outcomes are presented for each criterion. The degree of alignment expressed in the table is based on the combined judgments of the reviewers from the northern and southern workshops per content area.

As Table 2.15 demonstrates, alignment levels for both content areas were similar. For mathematics, the core items covered the breadth and depth of the content expectations in the standards to a very high degree. For English-language arts, the ELA reviewers found that the core items represented the breadth of those standards to a high degree, while the items matched the depth of the content standards to a modest degree.

**Table 2.15. Degree of Alignment Between Core CAHSEE Test Items and Relevant California Academic Content Standards for Math and ELA**

Content Area	Alignment Criteria			
	Categorical Concurrence	Depth of Knowledge Consistency	Range of Knowledge Correspondence	Balance of Representation
ELA	Highly Aligned	Partially Aligned	Highly Aligned	Highly Aligned
Math	Highly Aligned	Fully Aligned	Highly Aligned	Fully Aligned

***Item Review Workshops: Universal Test Design of the CAHSEE***

For the universal test design tasks, staff from the National Center for Educational Outcomes (NCEO) led the workshop participants in evaluating the February 2005 CAHSEE test form to ensure that the format, wording, and content of the tests are accessible to a wide variety of students. We provide a brief discussion of universal test design, as well as the role of NCEO in developing guidelines for acceptable universal test design principles, before turning to the results.

***Universal Test Design in the Environment and Education***

Ron Mace, a wheelchair user and architect, originally coined the term *universal design*. In the mid-1970s, Mace became frustrated with watching his colleagues design structures that later had to be retrofitted to meet the needs of diverse users. In citing the

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

need for creating structures from the beginning to be maximally accessible, Mace began advocating for structures that could meet the needs of wheelchair users, elderly people, children, and people with sensory disabilities that were, at the same time, easily accessible to non-disabled users.

The Center for Universal Design (1997), an architectural center housed at North Carolina State University, defined universal design as “the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design.” Currently, universal design is commonplace in structures and products. Such design improves the quality of structures and products for disabled and non-disabled populations alike.

Examples of universal design can be found everywhere. Curb cuts, originally designed to allow wheelchair users access to sidewalks, are now frequently used by parents who have babies in strollers, bicycle riders, and shoppers using carts. Likewise, closed captioning technology is now a legal requirement for all new television sets in the United States. This requirement was fought for and won by activists in the Deaf community. Currently, however, people with hearing impairments are only a fraction of those who use closed caption technology. Heath clubs, bars, people who watch television while their partner sleeps, and English learners all benefit from such technology.

Educators now also frequently use the term universal design to refer to classroom environments. The term *Universal Design for Learning* (UDL) employs technology and pedagogical practices such as differentiated instruction and individualized learning to make classrooms accessible to all learners. In terms of design, UDL does not mean that classrooms are “one size fits all.” Rather, UDL seeks to make classroom environments and instruction accessible to all students through flexible approaches to teaching.

Educators also use the term universal design to describe assessments that are fair and flexible (yet valid) for a wide variety of students. In 2002, NCEO synthesized research from a variety of fields to comprise a list of elements that best described what a “universally designed assessment” includes (Thompson, Johnstone, & Thurlow, 2002). NCEO’s original list of elements included the following:

1. Universally designed assessments should be designed for an inclusive population.
2. Universally designed assessments should have precisely defined constructs.
3. Universally designed assessments should have accessible, non-biased items.
4. Universally designed assessments should be amenable to accommodations.
5. Universally designed assessments should provide simple, clear, and intuitive instructions and procedures.
6. Universally designed assessments should contain language and print that are maximally readable and comprehensible.

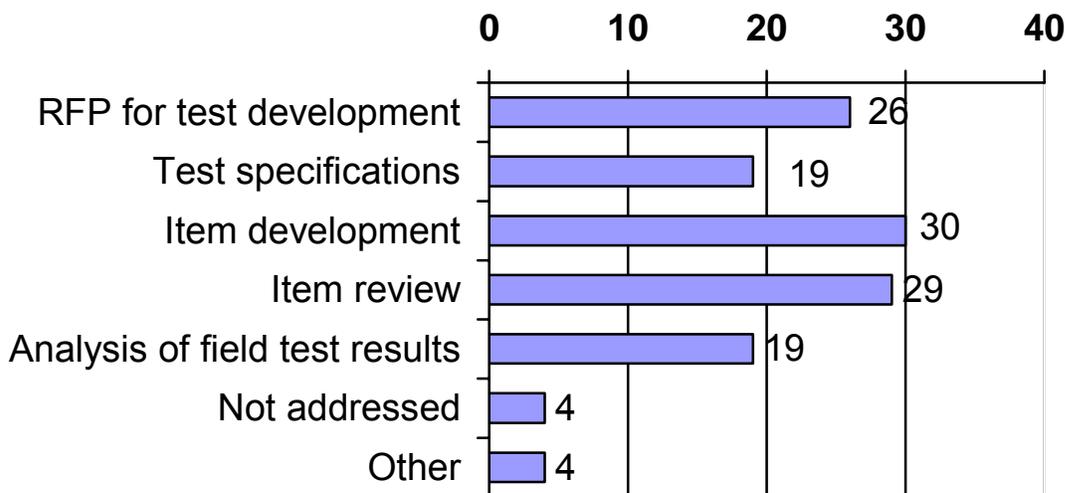
7. Universally designed assessments should have print and diagrams that are maximally legible.

### Research by NCEO and Other Organizations

In 2003, the United States Department of Education funded its first research study on universally designed assessments. From 2003 to 2005, NCEO, the Center for Applied Special Technology (CAST) and the University of Oregon each conducted research on improving accessibility of assessments for all students, including students with disabilities. As a result of this research and federal policy<sup>4</sup>, states have gradually become more amenable to the idea of universal design of assessments.

Currently, 26 states mention universal design in their requests for proposals from vendors, 19 states have universal design written into their test specifications, 30 states included universal design reviews in their item reviews, and 19 states analyzed field test results for possible design issues (Thompson, Johnstone, Thurlow, & Altman, 2005). Figure 1 (below) demonstrates the numbers of states that now include some form of universal design in their item reviews.

In response to the growing need for specific information about universal design, NCEO conducted a Delphi Study in an effort to validate Thompson et al.'s (2002) *Elements of Universally Designed Assessments* and to create a list of Considerations for Universally Designed Assessments that states could use to review items for potential design issues (Thompson, Johnstone, Anderson & Miller, 2005).



Source: Thompson, Johnstone, Thurlow, and Altman, 2005.

**Figure 2.1. Number of states that include universal design in test development.**

Thompson, Johnstone, Anderson, and Miller's 2005 *Considerations for Universally Designed Assessments* built on Thompson et al.'s 2002 *Elements* to create

<sup>4</sup> Assessment accessibility language is found in the No Child Left Behind Act of 2001 and "universal design" language is found in the Individuals with Disabilities Education Act of 2004.

## **Independent Evaluation of the CAHSEE: Third Biennial Report**

---

a list of issues to consider when reviewing items and tests. Experts from the fields of learning disabilities, English Language Learners, reading, mathematics, technology, and assessment discussed (on-line) the issues surrounding each of NCEO's considerations. The final product was a validated list of considerations that could be used by states when addressing universal test design issues. Although this list is not exhaustive, it provides a starting point for states to determine if the products they purchased from vendors act in accordance with universal design principles. The considerations finalized by NCEO's expert review panel included:

### ***Measure what it intends to measure***

- Reflect the intended content standards (reviewers have information about the content being measured).
- Minimize knowledge and skills required beyond what is intended for measurement.

### ***Respect the diversity of the assessment population***

- Be sensitive to test taker characteristics and experiences (consider gender, age, race/ethnicity, socio-economic level, region, disability, and language).
- Avoid content that might unfairly advantage or disadvantage any student subgroup.

### ***Have a clear format for text***

- Standard typeface
- Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group), and adaptable font size for computers
- High contrast between color of text and background
- Sufficient blank space (leading) between lines of text
- Staggered right margins (no right justification)

### ***Have clear visuals (when essential to item)***

- Use visuals when needed to answer the question.
- Use visuals with clearly defined features (minimum use of gray scale and shading).
- Ensure sufficient contrast between colors.
- Do not rely on color alone to convey important information or distinctions.
- Label visuals.

### ***Have concise and readable text***

- Keep to commonly used words (except vocabulary being tested).
- Use vocabulary appropriate for grade level.
- Avoid use of unnecessary words.

- Avoid idioms unless idiomatic speech is being measured.
- Avoid or define technical terms and abbreviations if not related to the content being measured.
- Use sentence complexity that is appropriate for grade level.
- Clearly identify the question to be answered.

***Allow changes to its format without changing its meaning or difficulty (including visual or memory load)***

- Allows for the use of Braille or other tactile format
- Allows for signing to a student
- Allows for the use of oral presentation to a student
- Allows for the use of assistive technology
- Allows for translation into another language

***Have an overall appearance that is clean and organized***

- All visuals (e.g., images, pictures) and text provide information necessary to respond to the item.
- Information is organized in a manner consistent with an academic English framework, with a left-right, top-bottom flow.
- Booklets/materials can be handled easily with limited motor coordination.
- Response formats are easily matched to question.
- The test includes space for student to take notes (on the screen for computer-based testing (CBT)) or extra white space with paper-pencil

An annotated list of the research supporting each of the considerations is found in Appendix D.

***Universal Test Design and the CAHSEE***

Prior to the evaluation study conducted by NCEO, the State of California and its vendor, ETS, had already expressed interest in ensuring that the CAHSEE was universally designed. California State educational law, section 60061.8 requires that educational endeavors (including assessment) must be universally designed. In response, ETS' project manager has conducted trainings with item designers about universal test design. All trainings were based on NCEO guidelines and other research related to accessibility of assessments.

***Universal Test Design Methods and Procedures***

***Procedures.*** The CAHSEE item reviews for the two workshops followed identical procedures. First, NCEO research staff trained reviewers to notice Considerations for Universal Design. Staff conducted training using a PowerPoint presentation that was also provided to reviewers as a handout. NCEO Universal Design staff provided

## **Independent Evaluation of the CAHSEE: Third Biennial Report**

---

information for reviewers and led discussion about universal design for approximately one hour.

Next, reviewers were split into two groups. One group was made up of English-language arts teachers (including special education teachers) and one group was made up of mathematics teachers (including special education teachers). In Sacramento, a school psychologist from the California School for the Blind moved between the two rooms in order to provide assistance on issues related to visual impairment. Likewise, in Los Angeles, a teacher from the California School for the Deaf supported both English and mathematics reviewers.

Using the Considerations for Universal Design forms, reviewers examined actual CAHSEE items and flagged any items they thought raised issues. For example, one teacher might have found a bias issue with a particular item while another found an issue with language complexity on another item. Reviewers marked issues they found as well as items they thought had features that appeared universally designed. For every item that appeared problematic, reviewers commented on what issue was present, noted whether they requested further review from a disability or culture expert, or called for further research to be conducted on particular item features. By calling for a further review or research, reviewers were identifying an aspect of an item that might be suspect, while recognizing their lack of expertise in making a definitive judgment. Reviewers also completed the *Considerations* process and paperwork for the entire test. Consequently, issues that appeared often or that were found related to the entire test, such as test formatting or font size, were recorded separately rather than recording the issue for every item that demonstrated that particular issue. Reviewers spent about two hours on individual review of the two tests made up 79 and 92 items, respectively.

At the end of the individual item review, reviewers engaged in discussion about items. As larger groups (English-language arts and mathematics), reviewers discussed each item's merits and shortcomings. In the end, reviewers agreed upon specific issues found in items. Likewise, the reviewers reached consensus on issues pertaining to the whole test. Consensus-making discussions were facilitated by NCEO research staff and lasted approximately two hours. Unlike in the consensus-making discussions about alignment, reviewers were not able to quantify issues related to items and the test because the issues they raised (if any) were qualitative issues.

Upon completion of subject-area reviews of tests, mathematics and language arts reviewers reconvened as a large group to discuss large group issues found across both tests (language arts and mathematics) and to evaluate the training and item review processes. This final discussion lasted approximately 20 minutes.

### ***Universal Test Design Analysis and Results***

This section of the report includes a summary of the item review results for mathematics and English-language arts. Several examples of results are included here to highlight reviewers' evaluations. These results represent the consensus ratings by

the group after individual review of items. Item-specific information is not included here due to test security concerns. Comments on specific test questions were provided to CDE and to the test developer.

Data were analyzed in the traditions of qualitative research, i.e., all data were examined and organized into large themes to produce meaningful information for readers. The following analyses took place in June and July of 2005:

1. Qualitative analysis of item-level data by subject area group (item-by-item analysis of consensus reports).
2. Quantitative analysis of whole-test issues by individuals (whole test issues raised by individuals).
3. Qualitative analysis of whole-test issues by subject area group (issues raised by whole groups regarding the whole test).

***Qualitative Results: Language Arts Item-by-Item Data as per Consensus by Language Arts Reviewers.*** Overall, reviewers found many ELA items to be well designed. They did, however, take issue with several items. According to our reviewers, only a few of these items had potentially *major* problems (i.e., significant enough issues to recommend that items be reexamined or removed from the test). According to reviewers, *major* problems were found in items that followed passages. In these items, reviewers were concerned that items might require students to have experiences that many students of low socioeconomic status did not have. Specifically, reviewers were concerned that items might advantage students of middle to high socioeconomic status because of the types of experiences referred to in the items. Likewise, reviewers found that, in some items with major issues, references to visual or auditory stimuli may have introduced bias against students who are blind or deaf. Most items that reviewers flagged, however, were considered to have potentially *minor* problems (i.e., minor changes were recommended but the overall item was deemed acceptable).

If corrected, the issues brought up by reviewers might improve the CAHSEE's overall design, readability, and accessibility. Specifically, only 11 items and 1 passage presented potentially major problems for reviewers. Several items and passages, however, were deemed to have potentially minor issues related to design. Among the categories that appeared to have the most minor problems for reviewers were diversity issues (11 items and 3 passages), readability issues (11 items, 1 passage, and 1 writing prompt), and formatting issues (32 items, 2 passages, and 1 writing prompt). Among these categories, reviewers most often questioned items' and passages' dependence on visual and auditory cues and reference to events that students of low socioeconomic status may not typically experience (diversity issues), the use of idiomatic or overly-complex language that was not imperative to the item's constructs (readability), and the lack of leading (white space) between lines of text (format issues).

In sum, 24 English-language arts items, but no English-language arts passages were considered to be problem-free. Reviewers found what might be minor problems

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

with 43 items and 2 passages, and what might be at least one major issue for 12 items and 5 passages.

**Qualitative Results: Mathematics Item-by-Item Data as per Consensus by Mathematics Reviewers.** As a whole, mathematics reviewers reached consensus quickly. These reviewers found many items to be well designed overall, but they did note minor issues with several items. Mathematics reviewers labeled only a few items as having potentially *major* problems, such as (a) an item that was worded in a manner that gave the answer away, (b) an item with two answer choices that could be potentially correct, (c) an item that did not align with standards, (d) an item with misleading visuals, and (e) an item that could cause confusion when presented under read-aloud accommodation conditions. Among the *minor* issues that could be addressed to improve the CAHSEE's overall design are issues related to readability and accessibility. The categories that appeared to have the most minor problems for reviewers were formatting issues (34 items), readability issues (24 items), and standards/assessment-related issues (12 items). Among these categories, reviewers were most often concerned that the graphs were too small (and graph grid lines did not have sufficient contrast), that equations were not given a separate line in the item to prevent confusion of signs, that equations were frequently written in sentence form rather than in numeric form (for example, the words "is equivalent to" were used instead of an "=" sign), that answer choices were arranged in a potentially confusing way on graph items, and that some items did not assess the intended standard. Only 4 mathematics items presented what might be major problems for reviewers.

In sum, reviewers found no problems at all with 28 items. Reviewers found potentially minor problems with 61 items, and what might be at least one major issue for 4 items.

**Quantitative Analysis of Whole-Test Issues by Individuals (Whole Test Issues Raised by Individuals).** After evaluating the individual ELA and math items, reviewers were asked to identify what they saw as themes (both strengths and weakness) in each content area. These themes, or whole-test issues, draw attention to common patterns that could be addressed. First, reviewers in each content area were asked to make independent judgments of the whole-test issues. Results from individual reviewers are reported below.

Table 2.16 lists the types of themes that emerged for English language arts and Table 2.17 lists the themes that emerged for mathematics. In both tables, Column 1 lists the broad themes that emerged, while Column 2 identifies specific sub-issues within these themes. Column 3 indicates the number of math reviewers who identified the issue. It should be noted that, if a reviewer identified one or more issues pertaining to the consideration (i.e., "Respects Diversity,") then the reviewer would be counted once for the consideration and then once for each sub-issue. For this reason, the number of reviewers listed next to each sub-issues will not typically equal the overall number for reviewers who identified broad areas of concern.

**Table 2.16. Individual, Whole Test Analysis of CAHSEE ELA Items (N = 14)**

Consideration	Sub-issue	Sub-issue Total	Consideration Total
Respects Diversity	Rural bias	4	7
	Vision bias	4	
	Hearing bias	3	
	SES bias	2	
	Autism bias	1	
	EL bias	1	
Concise and readable text	Low reading level	1	2
	High reading level	1	
	Directions ignorable	1	
Clear format	Response form color is confusing	1	5
	Inconsistent numbering pattern (i.e., up/down & left/right)	2	
	Writing prompt issues (i.e., two sets of instructions, skipped entirely, more space needed for planning)	2	
	Increase leading	2	
Clear visuals	Visuals are unclear/poor		4
	Distracting border		3
Amenable to accommodations			1
	Dictionaries should be allowed	1	1
Other	Essay points not clear	3	4
	Items do not always measure standards	2	

**Independent Evaluation of the CAHSEE: Third Biennial Report**

**Table 2.17. Individual, Whole Test Analysis of CAHSEE Math Items (N = 16)**

Consideration	Sub-issue	Sub-issue Total	Consideration Total
Respects Diversity	Vision bias	1	2
	Hearing bias	1	
Concise and readable text			11
	Simplify language	5	
	Minimize language	3	
	Maintain consistency in units between stem and response options	2	
	One equation per line	2	
	Keep prepositions attached to objects	1	
	Write out equations, not put in sentence	1	
	Word question consistently	1	
	Avoid proper names	1	
	Reading level too high on some items	1	
Clear format			10
	Increase space between items on page	4	
	Change format: A B above, C D below	3	
	Increase space between numbers	2	
	Increase leading	2	
	Enlarge font (esp. for exponents)	2	
Clear visuals			7
	Increase space around expressions	1	
	Enlarge grid	4	
	Increase contrast of grid lines & bars	3	
	Larger print	1	
Amenable to accommodations			0
	Lighen grid lines	1	
	Darken lines	1	
Other			9
	Test too long for one day	8	
	Lacks item type diversity (Only Multiple Choice)	3	
	Give graph paper	3	
	Give punch out ruler	2	
	Include math courses on answer form	2	
Shaded space between items on form	1		

In general, the majority of reviewers did not find whole-test issues with either the CAHSEE language arts or mathematics tests. The only exceptions included one consideration on the language-arts test (i.e., Respects Diversity) and two considerations on the math test (i.e., Concise and Readable Text, and Clear Format). Under Respects Diversity, reviewers reported that the language arts test included a disproportionate number of passages with content more familiar to students from rural areas, and a distinct lack of content relevant to students from urban areas. Additionally, reviewers expressed concern about the extent to which passage and subsequent items were biased against individuals with visual and hearing impairments.

On the mathematics test, Concise and Readable Text issues typically were related to the complexity of the vocabulary being used and item wordiness. Issues pertaining to Clear Format ranged from increasing workspace between items, changing the ordering of the items, and issues pertaining to visibility (e.g., increase line spacing, increase font size).

***Qualitative Analysis of Language Arts Whole-Test Issues by Subject Area Group (Issues Raised by Whole Groups Regarding the Whole Test).*** English-language arts reviewers also came to consensus on whole-test issues. These reviewers deemed the following features as potential problems with the CAHSEE test: (a) the test is too vested in multiple, long passages; (b) the directions for items and sections on the test are often poorly highlighted; (c) the passages appear biased against urban, low socioeconomic status students; (d) the visuals related to items were sometimes unclear and all visuals should have captions; (e) there was insufficient spacing between lines of text on items (leading); (f) passages contained many references that assumed experience with vision or hearing—such passages may be biased against students with visual or hearing disabilities; and (g) some of the language on the assessment was inconsistent with language used in state standards. Each of the language arts issues is presented below with a brief explanation.

***The test is too vested in multiple, long passages.*** Reviewers felt as if the test depended too heavily on reading passages that were very long. Reviewers found that there was a lack of variety in the length of passages. Reviewers agreed that some long passages were necessary in order to assess the reading proficiency of students, but expressed concern that too many long passages caused unnecessary cognitive demands.

***Directions were poorly highlighted.*** Reviewers pointed out several occasions where it was easy to ignore the directions provided because they were not visually highlighted. In these circumstances, reviewers were concerned that students may miss important information about an item or passage.

***Visuals were unclear, need captions.*** Reviewers argued that it was sometimes difficult to distinguish what the visuals placed next to passages portrayed. In many cases, reviewers argued that pictures were not clear enough to aid in comprehension. In addition, none of the visuals contained captions. Such captions are important for both students with visual impairments and students who may not have familiarity with the content of visuals.

***Insufficient line spacing between text in items.*** Although reviewers raised few complaints about the line spacing (leading) in passages, they expressed concern that text in items was insufficiently spaced (i.e., selected fonts resulted in letters spaced too close together). Reviewers commented that, on several items, text appeared jumbled because lines of text were too close. Although leading was sufficient on many items, it was inconsistent throughout the test.

## **Independent Evaluation of the CAHSEE: Third Biennial Report**

---

Passages assumed hearing or vision experience. Many of the passages in the CAHSEE alluded to sounds and sights as a way of describing the context of the story. Reviewers were concerned that dependence on such sensory imagery may cause difficulties for students who have sensory impairments. In this case, reviewers were concerned that students with hearing or visual disabilities would have difficulty accessing items.

Language used in assessment was inconsistent with that in standards. Many of the items asked students to refer to certain portions of passages or demonstrate certain skills. The instructions provided, however, often used terms that were not found in state standards. Reviewers were concerned with this inconsistency.

***Qualitative Analysis of Mathematics Whole-Test Issues by Subject Area Group (Issues Raised by Whole Groups Regarding the Whole Test).*** After completing individual rater whole-test item reviews, each group of content area reviewers came together to pinpoint the most important issues through a consensus discussion. The issues under discussion either recurred frequently in tests or were general design issues unrelated to particular items.

The mathematics reviewers deemed the following features as potential problems with the CAHSEE test: (a) the number of items per page (and related lack of space for students to take notes); (b) inconsistent leading and spacing between items; (c) the size and print contrast of graphs, (d) the presentation of equations, (e) the consistency of item stem and answers, and (f) the length of the test. Each of the mathematics issues is presented with an explanation below.

Items per page. Reviewers noted that items appeared cluttered on pages. The number of items per page was both visually challenging and gave students little room to take notes, calculate, etc. next to the items at the top of the page. Some reviewers suggested that the latter issue could be addressed by providing all students graph paper on which to work. Reviewers also suggested that when four items were presented per page, the items should be evenly spaced on the page to provide an equal amount of writing space for each one. As is, the top two items had little to no writing space and the bottom two items had sufficient (or more than enough) writing space.

Inconsistent line spacing and spacing between items. Reviewers noted that some items had sufficient line spacing. On others, however, they noted inconsistencies in the spacing between lines on items and in the spacing between letters on individual items throughout the test.

Size and print contrast of graphs. Reviewers noted that graphs were too small for some students to see. In addition, reviewers had issues with the lack of sharp contrast between the white and black grid lines on graphs. According to reviewers, these problems may cause students to misread data on graphs.

*Presentation of equations.* Reviewers noted that many equations were written as sentences. This, according to mathematics reviewers, was unnecessary. Rather, reviewers recommended that all equations be written in proper equation format. In addition, reviewers noted that when equations are written within a line of text they might be difficult to understand. Reviewers recommended that equations should be written on separate lines from all other text.

*Length of the test.* Finally, some reviewers were concerned that a test of 92 items was too demanding for a wide variety of students. These reviewers suggested that a shorter test could assess the same standards with fewer items.

### ***Universal Test Design Conclusions and Recommendations***

The purpose of the universal test design evaluation of the mathematics and English-language arts portions of the February 2005 CAHSEE test form was to determine whether these items are accessible to a wide range of students. Reviewers evaluated test items for format, organization, and content. The results from this investigation provide evidence in support of the efforts of the State of California to make the assessment system appropriate and accessible to all students required to take the CAHSEE.

The general conclusion is that most issues that reviewers found were deemed minor. In addition, many items were found to have no issues at all. For mathematics, the reviewers determined that many of the issues that arose centered on formatting (e.g., how equations were written; line spacing; and number of items per page). For ELA, the issues that arose dealt primarily with passages (e.g., passages appeared to favor the experience of middle-class, non-urban students without sensory impairments).

This study provides important information on how issues of universal test design can be assessed by content and population experts (i.e., teachers and other school personnel). The abundance of information found in the Results section provides a dataset that can be used for specific and targeted item-level test improvement.

Based on the findings in these investigations, we recommend that ETS review their current item development and review procedures against four goals to enhance the test design. These recommendations are based on data that emerged from the universal test design studies in June 2005 and universal design research over the past half-decade.

- 1. Ensure the CAHSEE has an inclusive test population.**

Although several items and passages appeared to present biased testing conditions against particular populations of students, this should not suggest that particular populations should be excluded from the CAHSEE. Studies such as the HumRRO alignment study and NCEO Universal Design study may improve the test for all students, including students with disabilities, English learners, and other students who traditionally underperform on standardized assessments. Excluding these populations while improvements

## Independent Evaluation of the CAHSEE: Third Biennial Report

---

- to the test are being made, however, could have serious implications for instruction. Therefore, as the CAHSEE progresses, all populations should take the test on its first administration. Data from field tests and actual administrations of the CAHSEE can then be used to make future decisions (see Recommendation No. 2).
2. **Cross-analyze item-level data.**

As noted above, reviewers found a variety of issues with individual items. Such information is a data source that should be seriously considered by stakeholders in the State of California. This information, however, represents only one data set. It is recommended that these data, combined with other data, be used to make decisions on items as assessments progress. For example, if reviewers determined that an item may be biased against a particular population of students and the field-test or live-test administration statistics also indicate bias, the item should be examined for change or omission from future tests. In addition to the data presented above and statistical analyses of items by population, the State of California may wish to conduct cognitive labs (think aloud studies) with particular populations of students for which reviewers found potentially problematic items. Such studies will provide another data set from which to make decisions. By combining the data from this study, statistical evidence, and cognitive lab studies, the State of California will have a triangulated data set from which to make item-level decisions. In the current data set, reviewers have raised red flags on particular item-level issues that should be taken as cautions for future analysis.
  3. **Changes to future CAHSEE tests should be made at the whole-test level first.**

Although reviewers found a variety of potential issues with individual items, reviewers also found that several issues appeared often, and therefore were considered whole test issues. Because of the repeated nature of the issues that arose as whole test issues, these should be considered for immediate change and correction. Many of the issues raised by reviewers are matters of simple changes in format (e.g., the spacing of mathematics items on each page, the placing of equations on separate lines of text, and the amount of leading between lines of text in items) and should be relatively inexpensive to make. Issues surrounding passages, however, may require more substantial investment. According to reviewers, passages that appear to advantage middle- to upper-class suburban students should not be completely removed from the CAHSEE. Rather, reviewers recommended that passages be more balanced to reflect the schema and experience base of the wide variety of students taking the CAHSEE (specifically mentioned were urban students, students with sensory impairments, and students of low socioeconomic status).
  4. **Revisit any issues related to alignment.**

Reviewers found few items that did not align to standards. Those that did not, however, should be revisited and revised as necessary. A test that is not well aligned to standards is not universally designed; therefore item-level data

from this study combined with HumRRO's study should provide the State of California with a succinct list of items to revise as necessary.

Overall, this study demonstrated that the State of California and its contractor, ETS, succeeded in creating a test without major design flaws. Those design issues that did arise should be addressed in prompt fashion, but a universal test design review of the CAHSEE, conducted by teachers, demonstrates that most items have only minor (if any) universal test design issues. The creation and improvement of any assessment is an ongoing and challenging process, but the willingness of the State of California and ETS to engage in alignment and universal test design studies early in the process (and as new versions are created) ensures that the CAHSEE will be in a constant state of improvement, will assess challenging standards, and will be accessible to all students.

### ***Item Review Workshop: Summary Findings***

The HumRRO item review workshops examined the quality and accessibility of the CAHSEE with California content experts. The studies assessed the February 2005 CAHSEE test form for alignment with the content standards and for appropriate format based on principles of universal test design.

The general conclusion from these investigations is positive. That is, the California educators involved in these workshops found the CAHSEE to be aligned with the content standards. Furthermore, these educators determined that the test is well constructed as a whole with mostly minor design issues.

Several specific recommendations follow from these findings. Concerning *alignment*, two recommendations are proposed:

1. Consider the definition and role of the mathematical reasoning standards. Assessment of these standards overlaps with the assessment of the more specific content standards and our reviewers had difficulties matching questions to these standards.
2. Consider creating a stronger match between the levels of cognitive complexity assessed by English-language arts items and those expected in the standards document for two standards: *Reading—Word Analysis, Fluency, and Systematic Vocabulary Development* and *Reading Comprehension (Focus on Informational Materials)*.

Recommendations for *universal test design* include reviewing test development and test form design procedures for the following goals:

1. Ensure the CAHSEE has an inclusive test population.
2. Cross-analyze item-level data.
3. Make changes to future CAHSEE tests at the whole-test level first.
4. Revisit any issues related to alignment.