# A Study of the Alignment of Florida's Sunshine State Standards with the Florida Comprehensive Assessment Test

## Reading

Conducted by the Learning Systems Institute

Laura Hassler, Ph.D.
Martha Beech, Ph.D
Karen DeMeester, Ph.D.

# Table of Contents

(Appendices are posted on the FCAT Web site at: http://fcat.fldoe.org/fcatpub5.asp.)

## Acknowledgments

This Alignment Study of Florida's Sunshine State Standards and the Florida Comprehensive Assessment Test was funded by the Florida Department of Education. In conducting this study we used Norman Webb's alignment criteria and process and his computerized Web Alignment Tool. We would like to thank Dr. Webb and Brian Vesperman for their assistance during this process.

We would also like to take this opportunity to thank the following alignment study group participants and especially the Group Leaders who provided training and guidance during the study:

**Language Arts**:

Group Leader: Mark Brunner, Ph.D., Coordinator of Elementary Education and Lake Region Schools, Citrus County School District
Max Hutto, Supervisor of Middle School Language Arts, Hillsborough County Schools
Fielding Hossley, Language Arts Chair (retired), Madison Middle School, Brevard County Schools
Kevin Smith, Reading Coach, Seminole High School, Seminole County School District
Teri Acquavita, Reading Curriculum Specialist K−5, Broward County School Board
Janet Langford, Elementary Education/ESOL Director, Gilchrist County School Board
Margarita D. Pinkos, Ed.D., Department of Multicultural Education, Palm Beach County School District

**Mathematics:**

Group Leader: Roberta Dilocker, Coordinator of Secondary Education (retired), Citrus County School District
Pat Wells, Grade 5 Chair of Mathematics, Clay County School District
Jean Giarrusso, Mathematics Resource Teacher, Department of Secondary, Adult and Community Education, Palm Beach County School District
Sandra M. Cook, Ph.D., Mathematics Teacher, Vernon High School, Washington County School District
Susan Saunders, Mathematics Teacher, Chipley High School, Washington County School District
Eileen Harris, Ph.D., Program Director, Area Center for Educational Enhancement (retired), FGCU
Carol McDaris, Mathematics Teacher, Fairview Middle School, Leon County School District

In addition, we would like to thank the following Florida Department of Education staff:

Cornelia Orr, Ph.D., Director, Assessment and School Performance
Kris Ellington, Director of K−12 Assessment, Assessment and School Performance
Bernard Stevens, Data Analysis, Reports, and Psychometric Services, Assessment and School Performance
Lyla Springfield, FCAT Reading, Elementary Specialist, Test Development Center
Donna Wolak, FCAT Reading Coordinator, Test Development Center
Vince Verges, FCAT Mathematics Coordinator, Test Development Center

# Executive Summary

The No Child Left Behind Act of 2001 requires states to have high-quality assessments that align with challenging academic standards. The Florida Department of Education contracted with the Learning Systems Institute (LSI) at Florida State University to conduct a study of the alignment between the Sunshine State Standards (SSS) and the Florida Comprehensive Assessment Test (FCAT) in Reading and Mathematics. The FCAT assessments reviewed in this study were selected from test administrations from 2003−2005. This report presents the findings from the study assessing the alignment between the SSS for Language Arts and the Reading FCAT for Grades 3, 8, and 10. Overall, the results indicate that the SSS and FCAT are generally aligned for all three grades but that alignment could be improved if the Reading FCAT tested a broader range of the academic content described in the SSS.

## Sunshine State Standards Included in This Study

The Sunshine State Standards for Language Arts include five Strands (referred to as Standards throughout this report): Reading; Writing; Listening, Viewing, and Speaking; Language; and Literature. Because not all of the content described in the SSS benchmarks can be tested in the limits of a single assessment, the advisory committee responsible for deciding the content to be covered by the Reading FCAT determined that the test would only assess student knowledge and skills in the areas of Reading and Literature. Therefore, as the Reading FCAT Test Item and Performance Task Specifications indicates, the test was not intended or designed to test all of the SSS for Language Arts.

One of the goals of an alignment study of a state's standards and assessments, however, is to ascertain the degree to which **all** the academic content that students are expected to master is tested by the state's assessments. Therefore, all of the SSS for Language Arts that are not tested by other assessments and that could be tested in a paper and pencil format such as the FCAT were included in this study—Standard A: Reading; Standard D: Language; and Standard E: Literature. Standard B: Writing is assessed on a separate FCAT specifically for writing, and the content of Standard C: Listening, Viewing, and Speaking cannot be assessed using a pencil and paper format like the FCAT. As the Reading FCAT was designed to assess only Reading and Literature content, it is not surprising that reviewers identified few test items assessing academic content from Standard D: Language, and, therefore, none of the alignment criteria for this standard were met.

## The Alignment Criteria and Process

A group of six reviewers with expertise in Language Arts standards and assessments (three from the elementary level, two from the middle-school level, and one from the high-school level) completed the study at FSU from October 19−21, 2005. Dr. Norman Webb's alignment process was used to conduct the study, and his Web Alignment Tool, an Internet-based tool, was used to generate statistical reports indicating the degree of alignment between the SSS and FCAT based on Webb's five criteria:

- Categorical Concurrence—the degree to which the same or consistent categories of content appear in the standards and assessments.
- Depth-of-Knowledge Consistency—the degree to which the knowledge elicited from students on the assessment is as complex as what students are expected to know and do according to the applicable standard.
- Range-of-Knowledge Consistency—the degree to which the span of knowledge that students need to answer assessment items correctly corresponds to the span of knowledge expected of students according to the applicable standard.
- Balance of Representation—the degree to which benchmarks that fall under a specific standard are given relatively equal emphasis on the assessment.
- Source of Challenge—the degree to which the primary difficulty of the assessment items is significantly related to students' knowledge and skill in the content area as represented in the standard. (Webb, 2005, pp. 3-4)

During the alignment study, reviewers provided the information the WAT would need to determine the degree of alignment on each of the five criteria. They began by assigning levels of cognitive complexity (1 for low complexity, 2 for moderate complexity, and 3 for high complexity) to each of the benchmarks included in the standards and to each FCAT test item. The level of complexity assigned to a benchmark indicates the content complexity associated with the knowledge and skills that students are expected to master, and the level of complexity for a test item indicates the cognitive demand associated with the tasks or thinking that a student must perform to answer the item correctly. Reviewers also assigned each test item to a primary benchmark (and up to two secondary benchmarks) that they thought best reflected the academic content being tested by that item. The data resulting from these activities were input into the WAT program, and the program generated reports indicating the degree of alignment for four of the criteria: Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Consistency, and Balance of Representation. At the same time reviewers assigned the level of cognitive complexity and the primary and secondary benchmarks to a test item, they also noted whether the item had a Source-of-Challenge problem.

## Performance Ratings for the Alignment Criteria

In the reports generated by the WAT, an acceptable level of alignment for a criterion is indicated by YES, a weak level of alignment is indicated by WEAK, and an unacceptable level of alignment is indicated by NO. Below are descriptions of the criteria used to rate the degree of alignment.

Categorical Concurrence. Reviewers provide the information necessary to determine whether the assessment measures content from each standard when they assign the test items to the benchmarks. A standard has an acceptable level of alignment for this criterion, if six or more test items are assigned to its benchmarks. A weak level of alignment exists if five to six items are assigned to a standard's benchmarks, and the degree of alignment is considered unacceptable if less than five items are assigned to a standard's benchmarks.

Depth-of-Knowledge Consistency. Reviewers provide the information necessary to determine whether the cognitive complexity of the test items aligns with the complexity of the knowledge and skills described in the standards when they assign the levels of cognitive complexity to the benchmarks and test items. Acceptable consistency in the level of complexity exists if 50% or more of the benchmarks are tested by items of a level of complexity equal to or greater than that of the benchmark. The alignment is weak if 40%–50% of the benchmarks are tested by items of an appropriate complexity, and the alignment is unacceptable if less than 50% of the benchmarks are targeted by items of appropriate complexity.

Range-of-Knowledge Consistency. Reviewers provide the information necessary to determine whether the full range of academic content described in the standards is tested on the assessment when they assign the test items to the benchmarks. To achieve an acceptable rating for this criterion, 50% or more of a standard's benchmarks had to be targeted by at least one test item. The criterion received a weak rating if 41%–49% of the benchmarks were targeted and an unacceptable rating if 40% or fewer benchmarks were targeted by at least one test item.

Balance of Representation. Reviewers provide the information necessary to determine whether the standards' academic content is emphasized equally on the assessment when they assign test items to the benchmarks. The WAT uses these assignments to compute a balance index for the standard that reflects the distribution of test items among the standard's benchmarks. To achieve an acceptable rating for this criterion, the standard must have a balance index of .7 or more. A balance index of .6–.7 indicates a weak rating for this criterion, and a balance index of .6 or less indicates an unacceptable rating.

## Results of the Studies

Grade 3 Alignment

The following table shows the results of the alignment study of the Grade 3 Reading FCAT and the SSS for Grades 3–5. Overall, the standards and assessment for this grade are aligned, but alignment could be improved if the assessment tested a broader range of the content described in the standards.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 3 Language Arts

| Standards | Alignment Criteria | | | |
|---|---|---|---|---|
| | Categorical Concurrence | Depth-of-Knowledge Consistency | Range-of-Knowledge Consistency | Balance of Representation |
| A – Reading | YES | YES | WEAK | WEAK |
| D – Language | NO | NO | NO | NO |
| E – Literature | YES | YES | NO | YES |

Standards A and E met the criteria for Categorical Concurrence and Depth-of-Knowledge Consistency. Standard A was rated WEAK, however, in both the Range-of-Knowledge Consistency and the Balance-of-Representation criteria. Reviewers assigned FCAT items to only 40% of the benchmarks included in Standard A, and of those benchmarks targeted, some were overrepresented while others were underrepresented. To raise the Range-of-Knowledge Consistency rating to an acceptable level, 2 additional benchmarks would need to be targeted by at least one FCAT item. To improve the Balance-of-Representation rating, test items targeting overrepresented benchmarks, such as LA.A.2.2.1, could be replaced by items targeting less represented benchmarks.

Standard E did not meet the Range-of-Knowledge Consistency criterion but did meet the Balance-of-Representation criterion. Reviewers assigned FCAT items to only 27% of the benchmarks under Standard E. To raise this Range-of-Knowledge Consistency rating to an acceptable level, 3 additional benchmarks would need to be targeted by at least one FCAT item.

Grade 8 Alignment

The following table shows the results of the alignment study of the Grade 8 Reading FCAT and the SSS for Grades 6–8. Overall, the standards and assessment for this grade are aligned, but alignment could be improved if the assessment tested a broader range of the content described in the standards.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 8 Language Arts

| Standards | Alignment Criteria | | | |
|---|---|---|---|---|
| | Categorical Concurrence | Depth-of-Knowledge Consistency | Range-of-Knowledge Consistency | Balance of Representation |
| A – Reading | YES | YES | WEAK | YES |
| D – Language | NO | NO | NO | NO |
| E – Literature | YES | YES | NO | YES |

Standards A and E met the criteria for Categorical Concurrence, Depth-of-Knowledge Consistency, and Balance-of-Representation. Standard A was rated WEAK in Range-of-Knowledge Consistency, and Standard E did not meet this criterion. To raise the Range-of-Knowledge Consistency rating for Standard A to an acceptable level, 2 additional benchmarks would need to be targeted by at least one FCAT item, and for Standard E to meet the Range-of-Knowledge Consistency criterion, 5 additional benchmarks would need to be targeted.

### Grade 10 Alignment

The following table shows the results of the alignment study of the Grade 10 Reading FCAT and the SSS for Grades 9–12. Overall, the standards and assessment for this grade are aligned, but alignment could be improved if the assessment tested a broader range of the content described in Standard E.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 10 Language Arts

| Standards | Alignment Criteria | | | |
|---|---|---|---|---|
| | Categorical Concurrence | Depth-of-Knowledge Consistency | Range-of-Knowledge Consistency | Balance of Representation |
| A – Reading | YES | YES | YES | YES |
| D – Language | NO | NO | NO | NO |
| E – Literature | YES | YES | NO | YES |

Standard A met all the criteria for proper alignment, and Standard E met all the criteria except for Range-of-Knowledge Consistency. In order for Standard E to meet the Range-of-Knowledge Consistency criterion fully, test items would have to be developed to target 5 additional benchmarks.

# Introduction

The No Child Left Behind Act of 2001 requires that states have high-quality academic assessments that align with challenging standards. According to the legislation, assessments that are properly aligned should (a) cover the full range of content specified in the standards; (b) measure both what students know and what students can do in relation to the content areas described in the standards; (c) reflect the same degree and pattern of emphasis as the standards; (d) be as demanding in terms of cognitive complexity and level of difficulty as the standards; and (e) yield results that represent all achievement levels specified in the standards.

In the *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001* (April, 2004), the U.S. Department of Education recommends that a state use an external organization to conduct a study to evaluate the degree of alignment between its assessments and its academic standards. In response to this recommendation, the Florida Department of Education contracted with the Learning Systems Institute (LSI) at Florida State University to conduct a study of the alignment between the Sunshine State Standards (SSS) and the Florida Comprehensive Assessment Test (FCAT) in Reading and Mathematics for grades representing elementary, middle, and high school.

To conduct the alignment study, LSI convened a group of fourteen teachers with expertise in assessments and standards (seven in the area of Language Arts, and seven in the area of Mathematics) from October 19–21, 2005. Two Group Leaders, one to facilitate the Language Arts study and one to facilitate the Mathematics study, provided information, resources, and training for the twelve reviewers and facilitated other group activities required in the alignment study process.

Each group consisted of participants representing all three grade levels. The Language Arts group consisted of three representatives from the elementary level, two from the middle-school level, and one from the high-school level. The group of Mathematics reviewers consisted of one representative from the elementary level, two from the middle-school level, and three from the high-school level. The intent of this heterogeneous design was for the group members to provide each other with the content knowledge and expertise needed to evaluate the benchmarks and test items from all three grade levels.

During the two-and-a-half-day study, each group of reviewers (six in the Language Arts group and six in the Mathematics group) reviewed FCAT tests selected from 2003–2005 test administrations for three grades and the SSS benchmarks established for the corresponding grade levels. The grades and subjects reviewed were Grade 3 Reading, Grade 5 Mathematics, Grade 7 Mathematics, Grade 8 Reading, Grade 9 Mathematics, and Grade 10 Reading. The elementary-level benchmarks and FCATs were reviewed on the first day of the study, and the middle-school and high-school level benchmarks and FCATs were reviewed on the second day.

LSI used Dr. Norman Webb's process for analyzing alignment and his Internet-based Web Alignment Tool (WAT) to conduct this study. The WAT automates the process of aligning state standards and assessments by capturing the information about the standards and assessments acquired during the alignment review process and generating statistical reports that reveal the degree of alignment based on five criteria:

- Categorical Concurrence—the degree to which the same or consistent categories of content appear in the standards and assessments.
- Depth-of-Knowledge Consistency—the degree to which the knowledge elicited from students on the assessment is as complex as what students are expected to know and do according to the applicable standard.
- Range-of-Knowledge Consistency—the degree to which the span of knowledge that students need to answer assessment items correctly corresponds to the span of knowledge expected of students according to the applicable standard.
- Balance of Representation—the degree to which objectives that fall under a specific standard are given relatively equal emphasis on the assessment.
- Source of Challenge—the degree to which the primary difficulty of the assessment items is significantly related to students' knowledge and skill in the content area as represented in the standard.

To prepare for the alignment study, information about the FCAT tests to be reviewed and the SSS standards and benchmarks for the grade levels covered by these tests was input into the WAT program. During the alignment study, reviewers did not analyze the alignment based on **each** of these five criteria. Instead, they participated in four activities, which primarily focused on the Depth-of-Knowledge Consistency criterion. The data resulting from these activities were input into the WAT program, and the program used the data to assess the degree of alignment on each of the five criteria.

The alignment study began with a brief introduction describing the purpose of the study, the participants' role as external reviewers, the activities they would be participating in, and how these activities would reveal the degree of alignment between Florida's standards and assessments. After this introduction, reviewers joined their content area groups (Language Arts or Mathematics), and the Group Leaders provided training to prepare the reviewers for the work they would do during the study. The training focused primarily on the three levels of cognitive complexity that Florida uses to describe the cognitive demand of the FCAT test items (low complexity—requires recall and recognition; moderate complexity—requires flexible thinking and possibly informal reasoning and problem-solving; high complexity—requires analysis and abstract reasoning). (See Appendix C: Cognitive Complexity Classification of FCAT SSS Test Items.) Reviewers were provided resources describing these levels of complexity, and they practiced assigning the different levels to sample test items and benchmarks.

During the study, reviewers assigned codes (referred to as *coding* in this report) corresponding to these levels of complexity (1 for low complexity, 2 for moderate complexity, and 3 for high complexity) to each benchmark and each FCAT test item. The level of complexity assigned to a benchmark indicates the content complexity associated with the knowledge and skills that students are expected to master, and the level of complexity for a test item indicates the cognitive demand associated with the tasks or thinking that a student must perform to answer the item correctly. Although these levels of complexity are primarily used to describe test items, in order for the WAT to determine if the benchmarks and assessments align on the Depth-of-Knowledge Consistency criterion, the benchmarks also had to be coded. For example, if a skill described in a benchmark requires analysis (level 3) and the FCAT item intended to test the student's proficiency with that skill only requires recall or recognition (level 1), there is a weakness in alignment. In this instance, the FCAT item does not measure whether the student has achieved the advanced level of knowledge or skill described in the benchmark, and, therefore, it does not provide full information regarding whether the state's expectations for student learning are being met.

After training was completed, the reviewers began the elementary-level study, the first of three studies they would complete (elementary, middle, and high school). For each study, the reviewers began by analyzing and assigning a level of cognitive complexity to each of the benchmarks for the grade level they were reviewing. Each reviewer input his or her codes into the WAT program using lap-top computers provided by LSI. Once all the reviewers had finished, the WAT generated a report showing each reviewer's codes for the benchmarks, and the Group Leaders used this report to identify benchmarks that reviewers had coded differently. The Group Leader then facilitated a consensus process to arrive at a single, agreed-upon set of codes for the benchmarks. The WAT used the consensus codes from each study to compare to the FCAT item codes to determine alignment on the Depth-of-Knowledge Consistency criterion. LSI staff input the consensus codes into the WAT while reviewers began the next step in the alignment process—coding the FCAT test items.

The reviewers coded the FCAT items using the three levels of cognitive complexity and assigned each item to a primary SSS benchmark (and up to two secondary benchmarks). For example, Grade 5 Mathematics FCAT items were assigned to Grade 3–5 benchmarks. The reviewers recorded their codes and benchmark assignments on coding forms, and LSI staff input the codes into the WAT. The groups concluded their studies with debriefing discussions in which they expressed their opinions regarding overall alignment for that grade-level FCAT and benchmarks. These four activities—coding the benchmarks, establishing a set of consensus codes, coding the FCAT items and assigning them to benchmarks, and participating in debriefing discussions—were repeated twice on the next day of the study: once for the middle-school level study and once for the high-school level study.

On the final day of the alignment study, the two groups came back together for an overall debriefing discussion. LSI staff, the reviewers, and the Group Leaders discussed the overall alignment between the SSS benchmarks and FCATs, offered suggestions for improving that alignment, and provided feedback regarding the alignment study process.

The participants agreed that the SSS and the FCATs were aligned but that alignment could be improved. In terms of improving the alignment, the primary recommendation was to clarify the language of the benchmarks and make them more specific to grade level expectations. Language used in the benchmarks, such as “understands,” was often too vague and ambiguous and made matching FCAT items to benchmarks more difficult. The reviewers suggested using the language related to Norman Webb’s Depth-of-Knowledge Consistency criterion or the FCAT Classification of Cognitive Complexity to revise the benchmarks.

When asked how they thought studying the alignment between standards and assessments could positively influence instruction, they said that teachers could incorporate the levels of cognitive complexity into their instruction and assessments and that staff development should be provided to help teachers do this. They thought the cognitive complexity model was the missing piece that could take instruction to a higher level. The reviewers also said that teachers have to resort to FCAT test-prep materials because they are not sure how to interpret the benchmarks.

In terms of improving the study process, the reviewers suggested that the study be extended to three days (completing one study per day) to provide more time to practice with FCATs that have been released to the public. They felt that discussion of these tests would provide the opportunity to learn from each other and to take advantage of the group members’ expertise across grade levels. They said that the distribution of participants across grade levels was very helpful. They also thought that more time available for coding the FCAT items and assigning them to benchmarks would be beneficial.

# Alignment Criteria Used for This Study

The degree of alignment between the SSS benchmarks and the FCATs was determined based on five criteria identified by Dr. Norman Webb of the Wisconsin Center for Education Research at the University of Wisconsin. The following descriptions of these criteria are taken from Dr. Webb's *Web Alignment Tool (WAT) Training Manual* (2005, pp. 110-114) and reprinted here with the permission of the author.

In terms of this study, the "objectives" that Dr. Webb refers to in these definitions are equivalent to the SSS "benchmarks." Furthermore, instead of Dr. Webb's four levels of depth of knowledge, the three levels of cognitive complexity—low complexity, moderate complexity, and high complexity— described in the Florida Department of Education's Cognitive Complexity Classification of FCAT SSS Test Items (Appendix C) were used to code the benchmarks and the test items. Therefore, instead of coding items as levels 1–4, reviewers coded them as levels 1–3.

## Categorical Concurrence

An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment, if both documents incorporate the same content. *The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content from each standard. The analysis assumed that the assessment had to have at least six items measuring content from a standard in order an acceptable level of categorical concurrence to exist between the standard and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors have to be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. This indicates that about 63% of the group would be consistently classified as masters or nonmasters if two equivalent test administrations were employed. The agreement coefficient would increase if the cutoff score were increased to one standard deviation from the mean to .77 and, with a cutoff score of 1.5 standard deviations from the mean, to .88. Usually states do not report student results by standards or require students to achieve a specified cutoff score on subscales related to a standard. If a state did do this, then the state would seek a higher agreement coefficient than .63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a standard, and as a basis for making some decisions about students' knowledge of that standard. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the standard, as judged in this analysis, at least 50% of targeted objectives are hit by items of the appropriate complexity. Fifty percent, a conservative cutoff point, is based on the assumption that a minimal passing score for any one standard of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding objectives. For example, assume an assessment included six items related to one standard and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding objectives, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one objective. Some leeway was used in this analysis on this criterion. If a standard had between 40% to 50% of items at or above the depth-of-knowledge levels of the objectives, then it was reported that the criterion was "weakly" met.

The justification above for the 50% cutoff point is based on the assumption that the standard is balanced. If the standard is not balanced, this reasoning does not apply. You could have a situation where a student passes the assessment that meets the DOK Consistency criterion without actually answering a single question at an appropriate DOK Level. Here is an example of why the DOK Consistency calculation must be considered in conjunction with Balance:

Assume an assessment included 6 items related to a given standard, and that these items specifically targeted 3 of the 5 objectives that fell under the standard. Consider two different cases.

The first case is that this standard is balanced—each of the 3 targeted objectives was hit by exactly 2 items. If 4 of the 6 items had DOK values lower than the objectives they targeted, then the depth-of-knowledge consistency score for this standard would be 33%—not high enough to be considered aligned.

The second case is that this standard is not balanced—1 of the 3 targeted objectives was hit by 4 items and the other 2 targeted objectives were only hit by 1 item each. Here, you could still have 4 of the 6 items with DOK values lower than the objective they targeted, just as in the first case. But if these 4 items all targeted the same objective, then the depth-of-knowledge consistency score would be 66%—indicating good alignment for this criterion!

## Range-of-Knowledge Consistency

For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of objectives within the standard with one related assessment item/activity. Fifty percent of the objectives for a standard had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a standard. This assumes that each objective for a standard should be given equal weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one objective, the requirement that assessment items need to be related to more than 50% of the objectives for a standard increases the likelihood that students will have to demonstrate knowledge on more than one objective per standard to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the objectives. However, any restriction on the number of items included on the test will place an upper limit on the number of objectives that can be assessed. Range-of-knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of standards and a large number of objectives. If 50% or more of the objectives for a standard had a corresponding assessment item, then the range-of-knowledge correspondence criterion was met. If 41% to 49% of the objectives for a standard had a corresponding assessment item, the criterion was "weakly" met.

## Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion only considers the number of objectives within a standard hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these objectives. *The balance-of-representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. This index only considers the objectives for a standard that have at least one hit—i.e., one related assessment item per objective. The index is computed by considering the difference in the proportion of objectives and the proportion of hits assigned to the objective. An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the objectives for the given standard. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the objectives hit. Depending on the number of objectives and the number of hits, a unimodal distribution (most items related to one objective and only one item related to each of the remaining objectives) has an index value of less than .5. A bimodal distribution has an index value of around .55 or .6. Index values of .7 or higher indicate that items/activities are distributed among all of the objectives at least to some degree (e.g., every objective has at least two items) and is used as the acceptable level on this criterion. Index values between .6 and .7 indicate the balance-of-representation criterion has only been "weakly" met.

*Note on the balance index*: The index formula for the balance criterion is $1 – (\sum|1/(O) – I_k/(H)|) / 2$, where $I_k$ is the number of items hit corresponding to objective k, O is the total number of objectives hit within the standard, and H is the total number of items hit within the standard. The balance index does not reflect how many objectives were hit within the given standard, but only how the hits were distributed across the objectives that *were* hit within the standard. For example, a standard where only one of its 20 objectives was hit would have a balance index of 1, although it would have a range of only 0.05 (1/20). This is why Range and Balance need to be considered together in order to obtain a well-rounded indication of how welldistributed the items are within a given standard. For instance, if every objective in this same standard was hit once, except one objective which was hit 20 times, this would give a range of 1 but a balance of 0.53.

Objectives A and C are not hit by items (so they are irrelevant for this calculation), Objectives B and D are each hit by one assessment item, and Objective E is hit by four items. Then this standard would have a balance index of 0.67, which would give a Balance of Representation alignment value of WEAK. (See Table 5.1a.) On the other hand, if the same objective was hit by items exactly the same way, except that Objective E was only hit by three items, then the standard would have a balance index of 0.73, which would give a Balance of Representation alignment value of YES. (See Table 5.1b.)

Table 5.1a
*An Example of a Weakly Balanced Standard*

| Standard N: | # of hits |
|---|---|
| Objective A | 0 |
| Objective B | 1 |
| Objective C | 0 |
| Objective D | 1 |
| Objective E | 4 |

| Balance Index: | 0.67 |
|---|---|
| Alignment: | WEAK |

Table 5.1b
*An Example of a Balanced Standard*

| Standard N: | # of hits |
|---|---|
| Objective A | 0 |
| Objective B | 1 |
| Objective C | 0 |
| Objective D | 1 |
| Objective E | 3 |

| | |
|---|---|
| Balance Index: | 0.73 |
| Alignment: | YES |

Source-of-Challenge Criterion

The Source-of-Challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted language arts skill, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a Source-of-Challenge problem. Such item characteristics may result in some students a) not answering an assessment item, b) answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed, or c) answering an assessment item correctly even though they do not possess the understanding and skills that the assessment administrators believe the item to be assessing.

# Findings for the Language Arts Alignment Study

## Standards and Benchmarks Included in the Language Arts Study

The Sunshine State Standards for Language Arts include five Strands (referred to as Standards throughout this report): Reading; Writing; Listening, Viewing, and Speaking; Language; and Literature. Because not all of the content described in the SSS benchmarks can be tested in the limits of a single assessment, the advisory committee responsible for deciding the content to be covered by the Reading FCAT determined that the test would only assess student knowledge and skills in the areas of Reading and Literature. Therefore, as the Reading FCAT Test Item and Performance Task Specifications indicates, the test was not intended or designed to test all of the SSS for Language Arts.

One of the goals of an alignment study of a state's standards and assessments, however, is to ascertain the degree to which **all** the academic content that students are expected to master is tested by the state's assessments. Therefore, all of the SSS for Language Arts that are not tested by other assessments and that could be tested in a paper and pencil format such as the FCAT were included in this study—Standard A: Reading; Standard D: Language; and Standard E: Literature. Standard B: Writing is assessed on a separate FCAT specifically for writing, and the content of Standard C: Listening, Viewing, and Speaking cannot be assessed using a pencil and paper format like the FCAT. As the Reading FCAT was designed to assess only Reading and Literature content, it is not surprising that reviewers identified few test items assessing academic content from Standard D: Language, and, therefore, none of the alignment criteria for this standard were met.

## Inclusion of Benchmarks LA.A.2.2.7 and LA.E.2.2.1 in the Grade 8 and Grade 10 Studies

In their analysis of Grade 8 Reading FCAT and Grade 10 Reading FCAT, the reviewers coded a number of items (for Grade 8, 9 items, and for Grade 10, 12 items) to the elementary-level benchmark LA.E.2.2.1— "Recognized cause-and-effect relationships in literary texts." They also coded items on the Grade 8 and 10 FCATs to the elementary-level benchmark LA.A.2.2.7—"Recognizes the use of comparison and contrast in a text." On the Grade 8 FCAT, 9 items were coded to this benchmark, and on the Grade 10 FCAT, 11 items were coded to it. Grade 8 and Grade 10 FCAT items that test academic content that students are expected to have mastered by Grade 5 could indicate a weakness in alignment.

In this instance, however, these benchmarks are intended to be cumulative and extend across all grades, and the FCAT test specifications indicate that the content of these benchmarks is tested at a more sophisticated level on the tests designed for the higher grades. According to Florida Department of Education staff,

When the Sunshine State Standards were first created, the Florida educators who

> participated in the Reading Content Advisory Meetings felt all students should be able to recognize Cause/Effect and Compare/Contrast beginning in the elementary grades and continuing through high school. The only difference in the benchmarks across the grades would be the 'depth of content' to which the comparison or causal relationship would appear in the reading text. (Donna Wolak, personal communication, November 3, 2005).

The educators' intention is implemented through the benchmark clarifications for LA.A.2.2.7 and LA.E.2.2.1 included in the Reading Test Item and Performance Task Specifications. The benchmark clarifications describe what students are expected to know and do regarding cause-and-effect and comparison-contrast relationships in texts at the different grade levels. The following table presents the benchmark clarifications from the Reading Test Item and Performance Task Specifications to show the benchmarks' progression in the depth of content across the grade levels reviewed for this alignment study.

Depth of Content Across Grade Levels for LA.A.2.2.7 and LA.E.2.2.1

| Grade Levels Reviewed in Alignment Study | Benchmark Clarification for LA.A.2.2.7 | Benchmark Clarification for LA.E.2.2.1 |
|---|---|---|
| Grade 3 | The student identifies no more than two similar or dissimilar elements within a test or identifies how elements are alike or different within a single text. (p. 44) | The student identifies cause-and-effect relationships, stated or strongly implied, in literary text or informational text. (p. 76) |
| Grade 8 | The student identifies similar or dissimilar elements within or across texts and/or explains in writing how elements are alike or different. (p. 34) | The student identifies or explains in writing cause-and-effect relationships within or across literary and/or informational texts. (p. 80) |
| Grade 10 | The student identifies or explains in writing when or how comparison and contrast are used within or across texts. (p. 26) | The student recognizes or explains in writing cause-and-effect relationships within or across informational and/or literary texts. (p. 71) |

To make alignment for these benchmarks more precise, additional benchmarks could be developed for grades 6–8 and grades 9–12 to reflect the increase in the depth of content that students are expected to demonstrate on the FCAT at these higher grades.

## Levels of Cognitive Complexity of the Benchmarks

The No Child Left Behind Act requires states to have challenging academic standards that hold all students in the state to a high level of academic achievement. In addition to identifying the knowledge and skills that students are expected to acquire at each grade level, Florida's Sunshine State Standards benchmarks also suggest the cognitive demand or degree of critical thinking that students need to apply to master the knowledge and skills described. The expectation that students demonstrate critical thinking is described in Goal 3, Standard 4, of the Florida System of School Improvement and Accountability: "Florida students use creative thinking skills to generate new ideas, make the best decisions, recognize and solve problems through reasoning, interpret symbolic data, and develop efficient techniques for lifelong learning" (Florida Department of Education, 2005, 1).

To evaluate the degree to which the benchmarks achieve this goal, reviewers in the alignment study assessed the benchmarks in terms of the level of complex thinking students are required to use to master the knowledge and skills described in the benchmarks. They coded the benchmarks with the same levels of cognitive complexity that they used to code the FCAT items: low, moderate, and high.

The following table indicates the levels of cognitive complexity that reviewers assigned to the Sunshine State Standards benchmarks for the grades included in this study.

Percent of Benchmarks by Levels of Cognitive Complexity for Each Grade
Florida Alignment Analysis for Language Arts

| Grade | Number of Benchmarks | Levels of Cognitive Complexity | Number of Benchmarks by Level | Percentage within Standard by Level |
|---|---|---|---|---|
| Grade 3 | 28 | 1<br>2<br>3 | 0<br>21<br>7 | 0<br>75<br>25 |
| Grade 8 | 36 | 1<br>2<br>3 | 2<br>22<br>12 | 6<br>61<br>33 |
| Grade 10 | 35 | 1<br>2<br>3 | 1<br>17<br>17 | 4<br>48<br>48 |

According to the reviewers' coding, the Language Arts benchmarks require primarily moderate levels of cognitive complexity with increasingly higher levels of demand as students advance into higher grade levels. For all three grade levels, the reviewers identified very few benchmarks that required low levels of cognitive complexity, and by Grade 10, almost 50% of the academic content that students are expected to master requires a high level of complex thought and advanced skill. In order to achieve alignment between a state's standards and assessments, benchmarks that require higher levels of cognitive demand should be tested by assessment items of at least equal cognitive complexity.

**Content Covered by the Reading FCAT**

The following table provides information regarding how much of the academic content described in the benchmarks is covered by the Reading FCATs for each of the grades studied.

Average Number of FCAT Items (Hits) Corresponding to Standards for Each Grade
Florida Alignment Analysis for Language Arts

| Standard | Grade 3 | | Grade 8 | | Grade 10 | |
|---|---|---|---|---|---|---|
| A – Reading | 36 | 78% | 40 | 83% | 48 | 84% |
| D – Language | 0 | 0% | 0 | 0% | 0 | 0% |
| E – Literature | 10 | 22% | 8 | 17% | 9 | 16% |

The information presented in the table indicates that as students advance from one grade level to the next, they are tested on an increasing amount of the academic content contained in Standard A, but on approximately the same amount for Standard E. As mentioned earlier, the Reading FCAT was not designed to test Standard D content.

**Alignment of Grade 3 Sunshine State Standards Benchmarks and FCAT**

The following table shows the results of the alignment study of Grade 3 Language Arts benchmarks and the Grade 3 Reading FCAT.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 3 Language Arts

| Standards | Alignment Criteria | | | |
|---|---|---|---|---|
| | Categorical Concurrence | Depth-of-Knowledge Consistency | Range-of-Knowledge Consistency | Balance of Representation |
| A – Reading | YES | YES | WEAK | WEAK |
| D – Language | NO | NO | NO | NO |
| E – Literature | YES | YES | NO | YES |

According to the results shown, overall alignment between the benchmarks and the FCAT for Grade 3 was acceptable in the areas of Categorical Concurrence and Depth-of-Knowledge Consistency. As discussed earlier, none of the benchmarks related to Standard D: Language were tested; therefore, none of the alignment criteria for this standard were met.

Standard A: Reading

Reviewers assigned FCAT items to only 40% of the benchmarks included in Standard A (less than the 50% required for the criteria to be met), so the Range-of-Knowledge Consistency criterion for Standard A was rated WEAK. Of the 12 benchmarks under Standard A, 4, on average, were targeted by test items (Appendix B, Table 3.3). To raise the Range-of-Knowledge Consistency rating to an acceptable level, 2 additional benchmarks would need to be targeted by at least one FCAT item.

Reviewers assigned no FCAT items to the following benchmarks, so adding test items targeting these benchmarks could improve the Range-of-Knowledge Consistency rating.

Benchmarks Not Represented on Grade 3 Reading FCAT

| Benchmarks Receiving No Hits (Consensus Level of Cognitive Complexity) | Content of Benchmarks |
|---|---|
| LA.A.1.2.1 (2) | Uses a table of contents, index, headings, captions, illustrations, and major words to anticipate or predict content and purpose of a reading selection. |
| LA.A.1.2.4 (2) | Clarifies understanding by rereading, self-correction, summarizing, checking other sources, and class or group discussion. |
| LA.A.2.2.3 (2) | Recognizes when a text is primarily intended to persuade. |
| LA.A.2.2.4 (2) | Identifies specific personal preferences relative to fiction and nonfiction reading. |
| LA.A.2.2.5 (3) | Reads and organizes information for a variety of purposes, including making a report, conducting interviews, taking a test, and performing an authentic task. |
| LA.A.2.2.6 (2) | Recognizes the difference between fact and opinion presented in a text. |

It is important to note that in order to maintain the acceptable rating for Depth-of-Knowledge Consistency, items developed to target unrepresented benchmarks would need to be at or above the consensus levels of cognitive complexity that reviewers assigned to these benchmarks. For example, an additional test item developed to target LA.A.2.2.6 would need to be at a moderate or high level of cognitive complexity.

Standard A also received a WEAK rating for the Balance-of-Representation criterion. The Balance Index for this standard was .62 (a .7 is required to meet this criterion, and an index between .6 and .7 is considered WEAK). Therefore, in addition to having too few benchmarks hit by test items, of the benchmarks that did receive hits, those hits were not distributed evenly. Benchmark LA.A.2.2.1 received the most hits (28), and LA.A.1.2.3 followed with 16 hits. Each of the other benchmarks that were targeted by test items received 4 or fewer hits (Appendix B, Table 3.11).

To improve the Balance-of-Representation and Range-of-Knowledge Consistency ratings, the number of FCAT items targeted to these overrepresented benchmarks could be reduced and items targeted to benchmarks that received no hits or fewer hits could be substituted. FCAT item 18, which was assigned to LA.A.2.2.1, could be a good candidate for replacement because reviewers thought it had a Source-of-Challenge problem due to confusing wording and a misleading stem. Furthermore, item 18 also had a level of complexity lower than the consensus code for LA.A.2.2.1. (According to Appendix B, Table 3.12, the benchmark was coded a 2, and the average code for item 18 was 1.67.) To maintain or improve the Depth-of-Knowledge Consistency rating, items with lower levels of complexity, such as 2, 37, and 40, would make the best candidates for replacement.

During their debriefing discussion, the reviewers indicated that LA.A.2.2.7 was underrepresented; however, even though adding additional items targeting this benchmark could improve the Balance-of-Representation rating, it would not improve the Range-of-Knowledge Consistency rating.

Standard E: Literature

Because reviewers assigned FCAT items to only 27% of the benchmarks under Standard E, the Range-of-Knowledge Consistency criterion for Standard E was not met. Of the 10 benchmarks for this standard, 3, on average, were targeted by FCAT items (Appendix B, Table 3.3). To raise this Range-of-Knowledge Consistency rating to an acceptable level, 3 additional benchmarks would need to be targeted by at least one FCAT item. Only 3 benchmarks under Standard E received hits: LA.E.1.2.2 (6 hits), LA.E.1.2.3 (6 hits), and LA.E.2.2.1 (13 hits). (See Appendix B, Table 3.11.) Reviewers assigned no FCAT items to the following benchmarks, so adding test items targeting these benchmarks could improve the Range-of-Knowledge Consistency rating.

Benchmarks Not Represented on Grade 3 Reading FCAT

| Benchmarks Receiving No Hits (Consensus Level of Cognitive Complexity) | Content of Benchmarks |
|---|---|
| LA.E.1.2.1 (2) | Identifies the distinguishing features among fiction, drama, and poetry and identifies the major characteristics of nonfiction. |
| LA.E.1.2.4 (3) | Knows that the attitudes and values that exist in a time period affect the works that are written during that time period. |
| LA.E.1.2.5 (3) | Identifies and uses literary terminology appropriate to the grade level, including symbol, theme, simile, alliteration, and assonance. |

| | |
|---|---|
| LA.E.2.2.2 (3) | Recognizes and explains the effects of language, such as sensory words, rhymes, and choice of vocabulary and story structure, such as patterns, used in children's texts. |
| LA.E.2.2.3 (3) | Responds to a work of literature by explaining how the motives of the characters or the causes of events compare with those in his or her own life. |
| LA.E.2.2.4 (2) | Identifies the major theme in a story or nonfiction text. |
| LA.E.2.2.5 (3) | Forms his or her own ideas about what has been read in a literary text and uses specific information from the text to support these ideas. |

Removing 3 items targeted to benchmark LA.E.2.2.1 and replacing these with items that target these unrepresented benchmarks could improve the Range-of-Knowledge Consistency rating while not compromising the acceptable Balance-of-Representation rating. Possible candidates for replacement would be items that have levels of complexity lower than the benchmark's consensus level of complexity. (LA.E.2.2.1 has a consensus level of 2.) Possible items to remove and replace are 15, 25, and 27 (Appendix B, Table 3.12). In order to maintain the acceptable rating for Depth-of-Knowledge Consistency, items developed to target the unrepresented benchmarks would need to be at or above the targeted benchmarks' consensus levels of cognitive complexity. As the above table indicates, 5 out of the 7 unrepresented benchmarks have a high level of complexity (3), so new test items targeting these benchmarks would also have to be at a high level of complexity.

### Alignment of Grade 8 Sunshine State Standards Benchmarks and FCAT

The following table shows the results of the alignment study of Grade 8 Language Arts benchmarks and Grade 8 Reading FCAT.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 8 Language Arts

| Standards | Alignment Criteria | | | |
|---|---|---|---|---|
| | Categorical Concurrence | Depth-of-Knowledge Consistency | Range-of-Knowledge Consistency | Balance of Representation |
| A – Reading | YES | YES | WEAK | YES |
| D – Language | NO | NO | NO | NO |
| E – Literature | YES | YES | NO | YES |

According to the results shown, overall alignment between the benchmarks and the FCAT for Grade 8 is acceptable in the areas of Categorical Concurrence and the Depth-of-Knowledge Consistency. As discussed earlier, none of the benchmarks related to Standard D: Language were tested; therefore, none of the alignment criteria for this standard were met.

Standard A: Reading

At this grade level, reviewers assigned FCAT items to only 42% of the benchmarks included in Standard A, so the Range-of-Knowledge Consistency rating is WEAK for Standard A. Of the 13 benchmarks, 5, on average, were targeted by test items. To raise the Range-of-Knowledge Consistency rating to an acceptable level, 2 additional benchmarks would need to be targeted by at least one FCAT item (Appendix B, Table 8.3).

Reviewers assigned no FCAT items to the following benchmarks, so adding test items targeting these benchmarks could improve the Range-of-Knowledge Consistency rating.

Benchmarks Not Represented on the Grade 8 Reading FCAT

| Benchmarks Receiving No Hits (Consensus Level of Cognitive Complexity) | Content of Benchmarks |
|---|---|
| LA.A.1.3.1 (2) | Uses background knowledge of the subject and text structure knowledge to make complex predictions about content, purpose, and organization of the reading selection. |
| LA.A.1.3.3 (2) | Demonstrates consistent and effective use of interpersonal and academic vocabularies in reading, writing, listening, and speaking. |
| LA.A.1.3.4 (2) | Uses strategies to clarify meaning, such as rereading, note taking, summarizing, outlining, and writing a grade-level-appropriate report. |
| LA.A.2.3.3 (2) | Recognizes logical, ethical, and emotional appeals in texts. |
| LA.A.2.3.4 (2) | Uses a variety of reading materials to develop personal preferences in reading. |
| LA.A.2.3.6 (2) | Uses a variety of reference materials, including indexes, magazines, newspapers, and journals, and tools, including card catalogs and computer catalogs to gather information for research topics. |

| | |
|---|---|
| LA.A.2.3.7 (3) | Synthesizes and separates collected information into useful components using a variety of techniques, such as source cards, note cards, spreadsheets, and outlines. |

In order to maintain the acceptable rating for Depth-of-Knowledge Consistency, however, items developed to target the content described in these benchmarks would need to be at or above the consensus levels of cognitive complexity reviewers assigned to the benchmarks. Furthermore, some of the benchmarks, specifically LA.A.2.3.6 and LA.A.2.3.7, might be more difficult than others to test with FCAT.

Of the benchmarks that were targeted by test items, the distribution of hits among the benchmarks is relatively even (indicated by the YES for Balance-of-Representation criterion). However, benchmarks LA.A.2.3.1 and LA.A.1.3.2 received a greater number of hits, and one item could be taken from each of these benchmarks without jeopardizing the Balance of Representation (the test items targeted to these benchmarks with the lowest levels of complexity are items 35, 46, 7, 36, and 38), and items could be added to target 2 of the unrepresented benchmarks to meet the Range-of-Knowledge Consistency criterion.

Standard E: Literature

Reviewers assigned FCAT items to only 17% of the benchmarks for this standard. Of the 13 benchmarks, 2, on average, were targeted by test items. For Standard E to meet the Range-of-Knowledge Consistency criterion, test items would have to be developed to target 5 additional benchmarks. The only benchmarks consistently hit by test items for this standard are LA.E.2.3.1 and LA.E.2.2.1.

Reviewers assigned no FCAT items to the following benchmarks, so adding test items targeting these benchmarks could improve the Range-of-Knowledge Consistency rating.

Benchmarks Not Represented on Grade 8 Reading FCAT

| Benchmarks Receiving No Hits (Consensus Level of Cognitive Complexity) | Content of Benchmarks |
|---|---|
| LA.E.1.3.1 (3) | Identifies the defining characteristics of classic literature, such as timelessness, deals with universal themes and experiences, and communicates across cultures. |
| LA.E.1.3.2 (2) | Recognizes complex elements of plot, including setting, character development, conflicts, and resolutions. |

| | |
|---|---|
| LA.E.1.3.3 (2) | Understands various elements of authors' craft appropriate at this grade level, including word choice, symbolism, figurative language, mood, irony, foreshadowing, flashback, persuasion techniques, and point of view in both fiction and nonfiction. |
| LA.E.1.3.4 (3) | Knows how mood or meaning is conveyed in poetry (e.g., word choice such as dialect, invented words, concrete or abstract terms, sensory or figurative language; use of sentence structure, line length, punctuation, and rhythm). |
| LA.E.2.3.2 (3) | Responds to a work of literature by interpreting selected phrases, sentences, or passages and applying the information to personal life. |
| LA.E.2.3.3 (2) | Knows that a literary text may elicit a wide variety of valid responses. |
| LA.E.2.3.4 (2) | Knows ways in which literature reflects the diverse voices of people from various backgrounds. |
| LA.E.2.3.6 (3) | Identifies specific questions of personal importance and seeks to answer them through literature. |
| LA.E.2.3.7 (2) | Identifies specific interests and the literature that will satisfy those interests. |
| LA.E.2.3.8 (2) | Knows how a literary selection can expand or enrich personal viewpoints or experiences. |

LA.E.2.3.1 was targeted by the greatest number of test items (14) (Appendix B, Table 8.11). The number of items related to LA.E.2.3.1 could be reduced and items targeting unrepresented benchmarks substituted. Items 7 and 37 would be good candidates for replacement because they have the lowest levels of cognitive complexity. Table 8.12 (Appendix B) reveals that the majority of test items targeting Standard E benchmarks are at or above the consensus level of cognitive complexity of the benchmarks; therefore, it is more difficult to select items to replace. In order to maintain the acceptable rating for Depth-of-Knowledge Consistency, items developed to target the unrepresented benchmarks would also need to be of a complexity level at or above that of the benchmarks. As the table indicates, new test items would need to be of moderate or high cognitive complexity.

## Alignment of Grade 10 Sunshine State Standards Benchmarks and FCAT

The following table shows the results of the alignment study of Grade 10 Language Arts benchmarks and Grade 10 Reading FCAT.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 10 Language Arts

| Standards | Alignment Criteria | | | |
|---|---|---|---|---|
| | Categorical Concurrence | Depth-of-Knowledge Consistency | Range-of-Knowledge Consistency | Balance of Representation |
| A – Reading | YES | YES | YES | YES |
| D – Language | NO | NO | NO | NO |
| E – Literature | YES | YES | NO | YES |

According to the results shown, overall alignment between the benchmarks and the FCAT for Grade 10 is acceptable in the areas of Categorical Concurrence, Depth-of-Knowledge Consistency, and Balance of Representation. As discussed earlier, none of the benchmarks related to Standard D: Language were tested; therefore, none of the alignment criteria for this standard were met.

At this grade level, the Range-of-Knowledge Consistency criterion is not met for Standard E: Literature. Reviewers assigned FCAT items to only 15% of the benchmarks included in Standard E. In order to meet this criterion fully, test items would have to be developed to target 5 additional benchmarks (Appendix B, Table 10.3). Of the 14 benchmarks, the reviewers consistently assigned FCAT items to only 2: LA.E.2.4.1 and LA.E.2.2.1 (Appendix B, Table 10.11).

Reviewers assigned no FCAT items to the following benchmarks, so adding test items targeting these benchmarks could improve the Range-of-Knowledge Consistency rating.

Benchmarks Not Represented on Grade 10 Reading FCAT

| Benchmarks Receiving No Hits (Consensus Level of Cognitive Complexity) | Content of Benchmarks |
|---|---|
| LA.E.1.4.1 (2) | Identifies the characteristics that distinguish literary forms. |
| LA.E.1.4.2 (2) | Understands why certain literary works are considered classics. |
| LA.E.1.4.3 (3) | Identifies universal themes prevalent in the literature of all cultures. |

| | |
|---|---|
| LA.E.1.4.4 (2) | Understands the characteristics of major types of drama. |
| LA.E. 1.4.5 (3) | Understands the different stylistic, thematic, and technical qualities present in the literature of different cultures and historical periods. |
| LA.E.2.4.2 (2) | Understands the relationships between and among elements of literature, including characters, plot, setting, tone, point of view, and theme. |
| LA.E.2.4.3 (3) | Analyzes poetry for the ways in which poets inspire the reader to share emotions, such as the use of imagery, personification, and figures of speech, including simile and metaphor; and the use of sound, such as rhyme, rhythm, repetition, and alliteration. |
| LA.E.2.4.4 (2) | Understands the use of images and sounds to elicit the reader's emotions in both fiction and nonfiction. |
| LA.E.2.4.5 (3) | Analyzes the relationships among author's style, literary form, and intended impact on the reader. |
| LA.E.2.4.6 (3) | Recognizes and explains those elements in texts that prompt a personal response, such as connections between one's own life and the characters, events, motives, and causes of conflict in texts. |
| LA.E.2.4.7 (3) | Examines a literary selection from several critical perspectives. |

LA.E.2.2.1 was targeted by the most test items, so the number of items assigned to this benchmark could be reduced without jeopardizing the acceptable Balance-of-Representation rating. Possible items assigned to LA.E.2.2.1 that could be replaced are 26, 39, 45, and 50. All of these items were assigned, on average, a level of cognitive complexity lower than the level of complexity assigned to the benchmark, so replacing these items should not compromise the acceptable rating for Depth-of-Knowledge Consistency. As the information in the table indicates, all of the unrepresented benchmarks under Standard E were assigned at least a moderate level of cognitive complexity, and 6 out of the 11 benchmarks were assigned a high level of cognitive complexity. Therefore, any new items developed to target those 6 benchmarks would also need to be at a high level of complexity in order to maintain an acceptable rating for Depth-of-Knowledge Consistency.

## Source of Challenge

An FCAT item may have a Source-of-Challenge problem if some students could answer the item correctly even though they do not possess the knowledge or skills the item is intended to test or could answer the item incorrectly even if they do possess such knowledge and skills. Cultural bias or specialized knowledge could be reasons for an item to have a Source-of-Challenge problem. Tables 3.5, 8.5, and 10.5 (Appendix B) show reviewers' comments regarding Source-of-Challenge problems for FCAT items analyzed in this study.

According to three reviewers, item 18 on the Grade 3 Reading FCAT had a Source-of-Challenge problem due to confusing wording and a misleading stem. The reviewers noted no Source-of-Challenge issues for the Grade 8 Reading FCAT. On the Grade 10 Reading FCAT, one reviewer indicated that item 24 could have a Source-of-Challenge problem because the student may know the answer without even reading the passage.

## Notes

As reviewers coded FCAT items, they had the opportunity to record their comments about specific test items. These comments can be found in Appendix B, Tables 3.7, 8.7, and 10.7. The tables also indicate how many reviewers commented on each test item; for example, if an item number is listed more than once, this means that more than one reviewer made a comment about that item. Each reviewer's comments are shown.

In general, the reviewers' comments concerned the grade-appropriateness of vocabulary, the degree of difficulty of distractors (answer choice options), and inconsistent or confusing wording. In their notes, some reviewers commented that some of the words used in the test items represented vocabulary above the grade level being tested. In their comments on the Grade 3 FCAT, reviewers indicated that *performance* in item 1, *chirped* in item 28, and *snuggle* in item 30 might be above a Grade 3 vocabulary level. In their comments on the Grade 8 FCAT, reviewers indicated that *projectiles* in item 2 and *commodity* in item 28 might be above a Grade 8 vocabulary level, and although they did not cite specific words, they indicated that the vocabulary in items 27 and 29 was above grade level. There were no comments indicating that the vocabulary on the Grade 10 FCAT was at an inappropriate grade level.

Reviewers also commented on the degree of difficulty of the distractors used in multiple-choice items. Commenting on the Grade 3 FCAT, one reviewer thought that the distractors made item 7 too easy and some reviewers thought that items 13, 14, 18, 26, and 38 were made more difficult because of the distractors. According to the Cognitive Complexity Classification of FCAT SSS Test Items, "The cognitive complexity of a multiple-choice item is generally NOT dependent on the distractors (answer choices). These answer choice options may affect the difficulty of the item, but not the complexity of the item" (Florida Department of Education, 2005, 1). LSI staff asked reviewers if they considered the degree of difficulty of the distractors in determining the level of cognitive complexity of the test items. Four of the reviewers said that YES they did consider the difficulty of the distractors and increased the level of cognitive complexity they assigned to an item if they felt the distractors made the item more difficult. One reviewer said that NO, he did not, and one reviewer failed to respond.

Other comments made by reviewers were (a) item 13 on the Grade 3 FCAT could be confusing because the chart says "in or near water" but the answer says "creek"; (b) item 18 on the Grade 3 FCAT was awkwardly worded; and (c) item 31 on the Grade 10 FCAT should include the beginning of the quote.

## General Comments Made by Reviewers

### Grade 3 Alignment Study

In their debriefing discussion for Grade 3, the reviewers said that they thought the alignment between the SSS benchmarks and the FCAT was acceptable and that the test items covered the content described in the benchmarks. They also indicated that the levels of cognitive complexity described in the benchmarks were generally aligned to those of the FCAT items. They said, however, that it was easier to determine the level of cognitive complexity of an item if they felt that they had accurately matched the item to the proper benchmark. In some cases, they thought it was difficult to assign test items to benchmarks and to differentiate between two benchmarks. They thought the ambiguous language of some of the benchmarks, such as "understands," made it more difficult to determine the level of cognitive complexity. They recommended aligning the wording of the benchmarks to the language used to describe the levels of cognitive complexity for the FCAT classifications.

### Grade 8 Alignment Study

In their debriefing discussion for Grade 8, the reviewers said that they thought the alignment between the SSS benchmarks and the FCAT was acceptable and that the test items covered the content described in the benchmarks. They commented, however, that LA.A.2.3.8 ("Checks the validity and accuracy of information obtained from research, in such ways as differentiating fact and opinion, identifying strong vs. weak arguments, and recognizing that personal values influence the conclusions an author draws") was possibly underrepresented on the test. They also commented that some benchmarks, such as LA.A.2.3.1 ("Determines the main idea or essential message in a text and identifies relevant details and facts and patterns of organization"), are too universal and include too much content relevant to what students should know and be able to do. They also said that there appeared to be more secondary benchmarks targeted by test items for this FCAT than the Grade 3 FCAT. They thought the range of levels of cognitive complexity of the FCAT items on the Grade 8 FCAT seemed appropriate. As with the Grade 3 FCAT, reviewers thought it was sometimes difficult to determine which benchmark(s) to assign test items to, and they thought additional training in coding items to the benchmarks would have been helpful.

### Grade 10 Alignment Study

In their debriefing discussion for Grade 10, the reviewers said that they thought the alignment between the SSS benchmarks and the FCAT was acceptable and that the test items covered the content described in the benchmarks. They also said that the passages used in the Grade 10 FCAT were more interesting than those used in the Grade 8 FCAT and that there appeared to be more secondary benchmarks targeted by test items on the Grade 10 FCAT than on the previous tests. They also felt there was a more equal distribution of test items across the benchmarks for this FCAT. For Standard A: Reading, study results affirm this impression, as both the Range-of-Knowledge Consistency and Balance-of-Representation ratings were adequate for this standard. Again, they indicated that they were concerned about matching the test items to the primary and secondary benchmarks and suggested providing time in the review process to study the benchmarks prior to coding. They thought a group review of items on FCATs that have been released to the public and group practice coding those items to the benchmarks would have been helpful because the reviewers/educators from the secondary level seemed to have a better sense of which items matched which benchmarks.

### Reliability among Reviewers

The WAT generates statistical measures for the reliability of reviewer coding (a) for the levels of cognitive complexity coded to test items and (b) for the standards and benchmarks assigned to test items. The following table shows the reliability measures for the Language Arts alignment study.

Reviewer Reliability

| Grade Level | Intraclass Correlation for FCAT Items | Pairwise Agreement for Standards | Pairwise Agreement for Benchmarks |
|---|---|---|---|
| Grade 3 | 0.7544 | 0.8111 | 0.6889 |
| Grade 8 | 0.7998 | 0.8096 | 0.6202 |
| Grade 10 | 0.8704 | 0.8644 | 0.5501 |

The intraclass correlation for the levels of complexity coded to the test items measures the percent of variance in the data that is caused by differences between the items rather than the differences between the reviewers. For example, if an intraclass correlation measure is .60, then 60% of the variance in the data is due to differences between the items, while 40% is due to differences among reviewers. The intraclass correlation is considered good if it is greater than 0.8 and adequate if it is greater than 0.7 (Webb, 2005, p.115). All of the studies had adequate correlation, and Grade 10 had good correlation.

The reviewers indicated that the most difficult aspect of the alignment study was assigning each test item to the appropriate benchmark(s). The pairwise agreement measures are possible indicators of the effect this difficulty might have had on the coding. Pairwise agreement for a test item is calculated using a pair of reviewers. The value is computed by identifying which of the two reviewers had the highest number of benchmarks assigned to the test item. For example, if Reviewer A identifies three benchmarks that are targeted by test item 16 and Reviewer B only identifies one, the number they agree on (1) is divided by the highest number of benchmarks assigned (3, assigned by Reviewer A) to get the pairwise agreement for that test item. To get the pairwise agreement for the benchmarks for the whole grade-level study, the pairwise agreement for benchmarks is averaged over all the assessment items (115).

The pairwise agreement measure is almost always lower than the intraclass correlation measure (116). Based on the values presented in the above table, the reviewers in this study had reasonable agreement regarding which standards and benchmarks test items were targeting. According to Norman Webb, one would expect to have agreement at approximately .9, so the .8 agreement indicates ambiguity in the standards and benchmarks and/or a weakness in the training provided during the study related to assigning test items to the standards and benchmarks (Norman Webb, personal communication, December 7, 2005).

# References

Florida Department of Education. (2005). *Cognitive complexity classification of FCAT SSS test items*. Tallahassee: Author.

Florida Department of Education. (2001). Reading test item and performance task specifications. Retrieved from http://fcat.fldoe.org/fcatis01.asp.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002). Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states. A study of the State Collaborative on Assessment & Student Standards (SCASS) Technical Issues in Large-Scale Assessment (TILSA). Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2005). *Web Alignment Tool (WAT): Training Manual*. Draft Version 1.1. Wisconsin Center for Education Research, Council of Chief State School Officers. Retrieved on September 15, 2005, from http://www.wcer.wisc.edu/wat/index.aspx.

# Appendix A

## Group Consensus Values for Language Arts Alignment Study

| | |
|---|---|
| **Grade 3** | **Table 3.13** |
| **Grade 8** | **Table 8.13** |
| **Grade 10** | **Table 10.13** |

(Appendices are posted on the FCAT Web site at: http://fcat.fldoe.org/fcatpub5.asp.)

# Appendix B

## Web Alignment Tool Tables

| | |
|---|---|
| **Grade 3** | **Tables 3.1-3.12** |
| **Grade 8** | **Tables 8.1-8.12** |
| **Grade 10** | **Tables 10.1-10.12** |

(Appendices are posted on the FCAT Web site at: http://fcat.fldoe.org/fcatpub5.asp.)

# Appendix C

## Florida Department of Education's

## Cognitive Complexity Classification of FCAT SSS Test Items

(Appendices are posted on the FCAT Web site at: http://fcat.fldoe.org/fcatpub5.asp.)