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

Mathematics

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, 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 Mathematics and the Mathematics FCAT for Grades 5, 7, and 9. Overall, the results indicate that the SSS and FCAT are generally aligned for all three grades but that alignment needs to be improved by raising the level of cognitive complexity of the test items to better reflect the cognitive complexity of the content described in the SSS and, to a lesser degree, testing a broader range of the content described in the standards.

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
The following describe the results of the alignment studies for Grades 5, 7, and 9.

Grade 5 Alignment
The following table shows the results of the alignment study of the Grade 5 Mathematics FCAT and the SSS for Grades 3–5.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 5 Mathematics

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Consistency	Balance of Representation
A – Number Sense, Concepts, and Operations	YES	YES	YES	YES
B – Measurement	YES	WEAK	YES	WEAK
C – Geometry and Spatial Sense	YES	YES	YES	YES
D – Algebraic Thinking	YES	YES	YES	YES
E - Data Analysis and Probability	YES	YES	YES	YES

All standards at this grade level met the Categorical Concurrence and Range-of-Knowledge Consistency criteria. Standards A, C, D, and E also met the criteria for Depth-of-Knowledge Consistency and Balance of Representation and therefore met all the criteria for acceptable alignment. Standard B, however, was rated WEAK in the Depth-of-Knowledge Consistency and Balance-of-Representation criteria.

The percentage of items at or above the consensus level of cognitive complexity assigned to Standard B was 42%, which means that a student could correctly answer approximately 7 of the 12 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. To achieve a YES rating for this criterion, approximately 2 new test items of a higher level of complexity could be added or approximately 1 item of a higher level of complexity could be substituted for an existing item of lower complexity. Another alternative would be to revise 1 item to raise its level of complexity.

Standard B was also WEAK in the Balance-of-Representation rating, which means that of the Standard B benchmarks targeted by test items, some benchmarks were overrepresented while others were underrepresented. To improve the rating for Balance

of Representation, items targeting overrepresented benchmarks could be replaced by items targeting other benchmarks.

Grade 7 Alignment

The following table shows the results of the alignment study of the Grade 7 Mathematics FCAT and the SSS for Grades 6–8.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 7 Mathematics

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Consistency	Balance of Representation
A – Number Sense, Concepts, and Operations	YES	YES	WEAK	WEAK
B – Measurement	YES	YES	YES	YES
C – Geometry and Spatial Sense	YES	YES	YES	YES
D – Algebraic Thinking	YES	NO	YES	YES
E – Data Analysis and Probability	YES	NO	YES	YES

All the standards at this grade level met the Categorical Concurrence criterion. Standards B and C also met the Depth-of-Knowledge Consistency, Range-of-Knowledge Consistency, and Balance-of-Representation criteria and therefore met all the criteria for acceptable alignment. Standard A met the Depth-of-Knowledge Consistency criterion but was rated WEAK on the Range-of-Knowledge Consistency and Balance-of-Representation criteria. To improve the ratings on these criteria, test items could be developed to target additional Standard A benchmarks and items targeting overrepresented benchmarks could be replaced by items targeting other benchmarks.

Standards D and E met the Range-of-Knowledge Consistency and Balance-of-Representation criteria but failed to meet the Depth-of-Knowledge Consistency criterion. The percentage of items at or above the consensus level of cognitive complexity assigned to Standard D was 36%, which means that a student could correctly answer approximately 8 out of the 13 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. To achieve a YES rating for this criterion, approximately 4 new test items of a higher level of complexity could be added or approximately 2 items of a higher level of complexity could be substituted for existing items of lower complexity. Another alternative would be to revise 2 items to raise their levels of complexity.

The percentage of items at or above the consensus level of cognitive complexity assigned to Standard E was only 40%, which means that a student could correctly answer approximately 4 out of the 7 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. To achieve a YES rating for this criterion, approximately 2 new test items of a higher level of complexity could be added or approximately 1 item of a higher level of complexity could be substituted for an existing item of lower complexity. Another alternative would be to revise 1 item to raise its level of complexity.

Grade 9 Alignment

The following table shows the results of the alignment study of the Grade 9 Mathematics FCAT and the SSS for Grades 9–12.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 9 Mathematics

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Consistency	Balance of Representation
A – Number Sense, Concepts, and Operations	YES	YES	NO	YES
B – Measurement	YES	YES	YES	YES
C – Geometry and Spatial Sense	YES	WEAK	YES	YES
D – Algebraic Thinking	YES	NO	YES	YES
E – Data Analysis and Probability	YES	NO	YES	YES

All the standards at this grade level met the Categorical Concurrence criterion, and Standard B met all the criteria for acceptable alignment. Standard A met the Depth-of-Knowledge Consistency and the Balance-of-Representation criteria but failed to meet the Range-of-Knowledge Consistency criterion. Only 40% of the benchmarks under this standard were tested on the Grade 9 Mathematics FCAT. In order to meet the Range-of-Knowledge Consistency criterion fully, test items would have to be developed to target 2 additional benchmarks.

Standard C met the Balance-of-Representation and Range-of-Knowledge Consistency criteria but was rated WEAK on the Depth-of-Knowledge Consistency criterion. The percentage of items at or above the consensus level of cognitive complexity assigned to Standard C was 47%, which means that a student could correctly answer approximately 7 out of the 13 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. To achieve a YES rating for this criterion, approximately 1 new test item of a high level of

complexity could be added or approximately 1 item of a high level of complexity could be substituted for an existing item of lower complexity. Another alternative would be to revise 1 item to raise its complexity level to high.

Standards D and E also met the Balance-of-Representation and Range-of-Knowledge Consistency criteria but failed to meet the Depth-of-Knowledge Consistency criterion. The percentage of items at or above the consensus level of cognitive complexity assigned to Standard D was 31%, which means that a student could correctly answer approximately 5 out of the 7 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. To achieve a YES rating for this criterion, approximately 3 new test items of a higher level of complexity could be added or approximately 2 items of a higher level of complexity could be substituted for existing items of lower complexity. Another alternative would be to revise 2 items to raise their levels of complexity.

The percentage of items at or above the consensus level of cognitive complexity assigned to Standard E was 24%, which means that a student could correctly answer approximately 7 out of the 9 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. To achieve a YES rating for this criterion, approximately 5 new test items of a high level of complexity could be added or approximately 2 items of a high level of complexity could be substituted for existing items of lower complexity. Another alternative would be to revise 2 items to raise their complexity levels to high.

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.

Noteon 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 Mathematics Alignment Study

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 Mathematics

Grade	Number of Benchmarks	Levels of Cognitive Complexity	Number of Benchmarks by Level	Percentage within Standard by Level
Grade 5	34	1 2 3	9 15 10	27 44 29
Grade 7	36	1 2 3	9 21 6	25 58 17
Grade 9	36	1 2 3	7 19 10	19 53 28

According to the reviewers' coding, the Mathematics benchmarks reflect primarily low and moderate levels of content complexity. Although one might expect to see increasingly higher levels of demand as students advance into higher grade levels, this does not appear to be the case for the Mathematics benchmarks. The levels of cognitive complexity expected of students drops slightly at the middle-school level and then returns to a level consistent with the elementary level. For all three grades, the number of

benchmarks with low and high levels of complexity stays relatively the same. To achieve alignment between the standards and assessments, assessment items should have complexity levels at least as high as the benchmarks they are testing.

Content Covered by the Mathematics FCAT

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

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

Standard	Grade 5		Grade 7		Grade 9	
A – Number Sense, Concepts, and Operations	21	31%	13	24%	10	20%
B – Measurement	12	17%	11	20%	11	21%
C – Geometry and Spatial Sense	12	17%	9	17%	13	25%
D – Algebraic Thinking	9	13%	13	24%	8	16%
E – Data Analysis and Probability	15	22%	8	15%	9	18%

According to the information presented in the table, content related to Number Sense, Concepts, and Operations is most represented on the elementary-level assessment. The other standards are represented more evenly across the grade levels, with Geometry and Spatial Sense slightly more prominent on the high-school level test and Algebraic Thinking more prominent on the middle-school level test.

Alignment of Grade 5 Sunshine State Standards Benchmarks and FCAT

The following table shows the results of the alignment study for Grade 5 Mathematics.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 5 Mathematics

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Consistency	Balance of Representation
A – Number Sense, Concepts, and Operations	YES	YES	YES	YES
B – Measurement	YES	WEAK	YES	WEAK
C – Geometry and Spatial Sense	YES	YES	YES	YES
D – Algebraic Thinking	YES	YES	YES	YES
E - Data Analysis and Probability	YES	YES	YES	YES

According to the results shown, all the standards met the Categorical Concurrence and Range-of-Knowledge Consistency criteria, but Standard B was WEAK in the Depth-of-Knowledge and Balance-of-Representation criteria. The percentage of items at or above the consensus level of cognitive complexity assigned to Standard B was only 42%, which means that a student could correctly answer approximately 7 of the 12 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. The Depth-of-Knowledge Consistency rating can be improved by either adding one or two additional test items that have levels of cognitive complexity at least as high or higher than the benchmarks they are intended to target, or by replacing the less demanding items with items at least as demanding as the Standard B benchmarks. The average consensus level of complexity for Standard B is 2 (moderate) (Appendix A, Table 5.13). To achieve a YES rating for this criterion, approximately 2 new test items of a higher level of complexity could be added or approximately 1 item of a higher level of complexity could be substituted for an existing item of lower complexity. Another alternative would be to revise 1 item to raise its level of complexity.

Standard B was also WEAK in the Balance-of-Representation rating, which means that of the Standard B benchmarks targeted by test items, some benchmarks received a disproportionate share of those hits. Consequently, some benchmarks targeted on the test were overrepresented on the Grade 5 Mathematics FCAT while others were underrepresented. According to Table 5.11 (Appendix B), MA.B.1.2.1 and MA.B.1.2.2 received the most hits, so to improve the rating for Balance of Representation, items targeting these benchmarks could be replaced by items targeting other benchmarks. Table 5.11 also indicates that reviewers thought that test items were often testing the academic content of both MA.B.1.2.1 and MA.B.1.2.2. Both benchmarks were targeted by 11 items, and 9 of those items were coded to the same benchmark (items 1, 14, 20, 30, 34, 43, 44, 55, and 57).

When attempting to improve the Balance-of-Representation rating by replacing test items and selecting additional benchmarks to target, the WEAK rating for the Depth-of-Knowledge Consistency should also be addressed. The items replaced should be those with low levels of cognitive complexity, and the items replacing them should be of moderate to high levels of cognitive complexity. Whether the level should be moderate or high depends on the benchmark being targeted. The goal is to have the test item's level of complexity at or above the consensus level of complexity assigned to that benchmark during the study. (See Appendix A, Table 5.13 for the consensus levels for each of the standards and benchmarks.) The consensus level of cognitive complexity for MA.B.1.2.1 is a 3 (high), and all of the test items assigned to it have lower levels of complexity (5 items are low, and 6 are moderate). The consensus level of cognitive complexity for MA.B.1.2.2 is a 2 (moderate), and 7 of the test items assigned to it have low levels of complexity while 4 are at the same level (moderate). Possible items targeted to benchmarks MA.B.1.2.1 and MA.B.1.2.2 that could be replaced by items of a moderate or high level of complexity are 30, 34, 44, and 57 (Appendix B, Table 5.12).

Alignment of Grade 7 Sunshine State Standards Benchmarks and FCAT

The following table shows the results of the alignment study for Grade 7 Mathematics.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 7 Mathematics

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Consistency	Balance of Representation
A – Number Sense, Concepts, and Operations	YES	YES	WEAK	WEAK
B – Measurement	YES	YES	YES	YES
C – Geometry and Spatial Sense	YES	YES	YES	YES
D – Algebraic Thinking	YES	NO	YES	YES
E – Data Analysis and Probability	YES	NO	YES	YES

According to the results shown, Standard B and Standard C met all the criteria for proper alignment, but Standard A was rated WEAK in the Range-of-Knowledge Consistency and Balance-of-Representation criteria, and Standard D and Standard E failed to meet the Depth-of-Knowledge Consistency criterion.

Standard A: Number Sense, Concepts, and Operations

Based on the results of the study, Standard A did not have enough of its benchmarks targeted by test items, and of the benchmarks targeted by test items, some benchmarks received a disproportionate share of those hits. Consequently, some benchmarks targeted on the test were overrepresented while others were underrepresented. Table 7.3 (Appendix B) shows that the mean percentage of benchmarks (objectives) targeted (hit) by test items was 50%. A WEAK rating for the Range-of-Knowledge Consistency criterion is 40%–50%, while a YES rating is 50% or higher (Webb, 2005, p. 153); therefore, the criterion is almost fully met.

In terms of the Balance-of-Representation rating, Table 7.3 (Appendix B) indicates that, like the Range-of-Knowledge Consistency criterion, this criterion is almost fully met. Table 7.11 (Appendix B) indicates that MA.A.3.3.3 and MA.A.3.3.2 received the highest number of hits (13 and 11, respectively). Therefore, to improve the Balance-of-Representation rating items targeted to these benchmarks could be replaced by items targeting other benchmarks that were not targeted, such as MA.A.2.3.2, or that only received a few hits, such as MA.A.1.3.1, MA.A.2.3.1, or MA.A.5.3.1. When selecting items targeted to benchmarks MA.A.3.3.3 and MA.A.3.3.2 that could be replaced, the

levels of cognitive complexity of those items (as well as those of the new items) need to be considered in order to maintain the acceptable alignment for the Depth-of-Knowledge criterion. Possible items to replace would be 17, 18, or 21 (Appendix B, Table 7.12). Another strategy would be to replace an item coded to another standard that is overemphasized on the test (Norman Webb, personal communication, December 7, 2005).

Standard D: Algebraic Thinking

According to Table 7.2 (Appendix B), the percentage of items at or above the consensus level of cognitive complexity assigned to Standard D was only 36%, which means that a student could correctly answer approximately 8 out of the 13 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. The Depth-of-Knowledge Consistency rating can be improved by either adding additional test items that have levels of cognitive complexity at least as high or higher than the Standard D benchmarks they are intended to target, or by replacing the less demanding items with items at least as demanding as the Standard D benchmarks. The average consensus level of complexity for Standard D is 2 (moderate) (Appendix A, Table 7.13). To achieve a YES rating for this criterion, approximately 4 new test items of a higher level of complexity could be added or approximately 2 items of a higher level of complexity could be substituted for existing items of lower complexity. Another alternative would be to revise 2 items to raise their levels of complexity. Possible items to replace would be 10, 17, 18, 21, or 30 (Norman Webb, personal communication, December 7, 2005).

Standard E: Data Analysis and Probability

According to Table 7.2 (Appendix B), the percentage of items at or above the consensus level of cognitive complexity assigned to Standard E was only 40%, which means that a student could correctly answer approximately 4 out of the 7 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. The Depth-of-Knowledge Consistency rating can be improved by either adding additional test items that have levels of cognitive complexity at least as high or higher than the Standard E benchmarks they are intended to target, or by replacing the less demanding items with items at least as demanding as the Standard E benchmarks. The average consensus level of complexity for Standard E is 2 (moderate) (Appendix A, Table 7.13). To achieve a YES rating for this criterion, approximately 2 new test items of a higher level of complexity could be added or approximately 1 item of a higher level of complexity could be substituted for an existing item of lower complexity. Another alternative would be to revise 1 item to raise its level of complexity. Possible items to replace would be 28, 40, or 50 (Norman Webb, personal communication, December 7, 2005).

Alignment of Grade 9 Sunshine State Standards Benchmarks and FCAT

The following table shows the results of the alignment study for Grade 9 Mathematics.

Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
Florida Grade 9 Mathematics

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Consistency	Balance of Representation
A – Number Sense, Concepts, and Operations	YES	YES	NO	YES
B – Measurement	YES	YES	YES	YES
C – Geometry and Spatial Sense	YES	WEAK	YES	YES
D – Algebraic Thinking	YES	NO	YES	YES
E – Data Analysis and Probability	YES	NO	YES	YES

According to the results shown, Standard B fully met all the criteria for proper alignment. Standard A fully met all the criteria for proper alignment except for the Range-of-Knowledge Consistency criterion, and Standards C, D, and E fully met all the criteria for proper alignment except for the Depth-of-Knowledge Consistency criterion.

Standard A: Number Sense, Concepts, and Operations

Based on the results of the study, Standard A did not have enough of its benchmarks targeted by test items. Only 40% of the benchmarks under this standard were tested on the Grade 9 Mathematics FCAT (Appendix B, Table 9.3). In order to meet the Range-of-Knowledge Consistency fully, test items would have to be developed to target 2 additional benchmarks (Appendix B, Table 9.3). According to Table 9.11 (Appendix B), the following benchmarks were not targeted by any of the items on the Grade 9 Mathematics FCAT.

Benchmarks Not Represented on the Grade 9 Mathematics FCAT

Benchmarks Receiving No Hits (Consensus Level of Cognitive Complexity)	Content of Benchmarks
MA.A.1.4.1 (1)	Associates verbal names, written word names, and standard numerals with integers, rational numbers, irrational numbers, real numbers, and complex numbers.
MA.A.2.4.2 (1)	Understands and uses the real number system.

Benchmarks Receiving No Hits (Consensus Level of Cognitive Complexity)	Content of Benchmarks
MA.A.2.4.3 (1)	Understands the structure of the complex number system.
MA.A.3.4.1 (3)	Understands and explains the effects of addition, subtraction, multiplication, and division on real numbers, including square roots, exponents, and appropriate inverse relationships.

To improve the Range-of-Knowledge Consistency rating, additional test items would need to be developed to target these benchmarks or existing items targeting other benchmarks (especially those that received the greatest number of hits, such as MA.A.3.4.3, so that the acceptable Balance of Representation would not be jeopardized) could be replaced by items targeting the unrepresented benchmarks.

Another consideration in replacing test items is the level of cognitive complexity of those items and the benchmarks they are intended to target. As the table above reveals, the content of 3 of the unrepresented benchmarks was considered by reviewers to be of low cognitive complexity, so adding items to test this content, particularly at the ninth-grade level, may not be desirable. Developing an additional item to target benchmark MA.A.3.4.1, as long as the item is of a high level of cognitive complexity like the benchmark, would elevate the complexity level of the test and might be a more desirable approach.

Standard C: Geometry and Spatial Sense

According to Table 9.2 (Appendix B), the percentage of items at or above the consensus level of cognitive complexity assigned to Standard C was 47%, which means that a student could correctly answer approximately 7 out of the 13 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. The Depth-of-Knowledge Consistency rating can be improved by either adding additional test items that have levels of cognitive complexity at least as high or higher than the Standard C benchmarks they are intended to target, or by replacing the less demanding items with items at least as demanding as the Standard C benchmarks. The average consensus level of complexity for Standard C is 3 (high) (Appendix A, Table 9.13). To achieve a YES rating for this criterion, approximately 1 new test item of a high level of complexity could be added or approximately 1 item of a high level of complexity could be substituted for an existing item of lower complexity. Another alternative would be to revise 1 item to raise its complexity level to high.

Standard D: Algebraic Thinking

According to Table 9.2 (Appendix B), the percentage of items at or above the consensus level of cognitive complexity assigned to Standard D was 31%, which means that a student could correctly answer approximately 5 out of the 7 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. The Depth-of-Knowledge Consistency rating can be improved by either adding additional test items that have levels of cognitive complexity at least as high or higher than the Standard D benchmarks they are intended to target, or by replacing the less demanding items with items at least as demanding as the Standard D benchmarks. The average consensus level of complexity for Standard D is 2 (moderate) (Appendix A, Table 9.13). To achieve a YES rating for this criterion, approximately 3 new test items of a higher level of complexity could be added or approximately 2 items of a higher level of complexity could be substituted for existing items of lower complexity. Another alternative would be to revise 2 items to raise their levels of complexity.

Standard E: Data Analysis and Probability

According to Table 9.2 (Appendix B), the percentage of items at or above the consensus level of cognitive complexity assigned to Standard E was 24%, which means that a student could correctly answer approximately 7 out of the 9 test items targeted to this standard without ever answering an item with as high a cognitive complexity as the knowledge and skills described in the standard. The Depth-of-Knowledge Consistency rating can be improved by either adding additional test items that have levels of cognitive complexity at least as high or higher than the Standard E benchmarks they are intended to target, or by replacing the less demanding items with items at least as demanding as the Standard E benchmarks. The average consensus level of complexity for Standard E is 3 (high) (Appendix A, Table 9.13). To achieve a YES rating for this criterion, approximately 5 new test items of a high level of complexity could be added or approximately 2 items of a high level of complexity could be substituted for existing items of lower complexity. Another alternative would be to revise 2 items to raise their complexity levels to high.

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. Appendix B, Tables 5.5, 7.5, and 9.5 show reviewers' comments regarding Source-of-Challenge problems for FCAT items analyzed in this study.

The reviewers noted no Source-of-Challenge problems for the Grade 5 FCAT and the Grade 7 FCAT. According to one reviewer, items 23 and 45 on the Grade 9 FCAT could

present a Source-of-Challenge problem because item 45 gives the formula needed to solve problem 23. During the debriefing discussion for the Grade 9 FCAT, the other reviewers agreed that this could be a Source-of-Challenge problem, even though they did not note it themselves.

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 5.7, 7.7, and 9.7. 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.

The following comments were made by reviewers in reference to the Grade 5 FCAT. One reviewer thought that item 11 targeted the least difficult aspects of MA.E.1.2.1. One reviewer thought that item 15 was more difficult because it required two steps, and one reviewer thought that item 20 was more difficult because the student had to change hours into minutes to answer the question. One reviewer thought that item 24 was "nonroutine." One reviewer thought that the use of pattern in item 34 could be confusing, and one reviewer thought that item 43 was difficult because it required estimation.

The following comments were made by reviewers in reference to the Grade 7 FCAT. One reviewer thought that item 5 was too easy for Grade 7, and one reviewer thought that item 52 was too easy. One reviewer thought that there were too many problems like item 18. Some reviewers noted items they thought didn't really fit any standard/benchmark very well: item 27 (two reviewers), item 28 (two reviewers), item 36 (two reviewers), and item 49 (two reviewers).

Some of the comments made regarding the Grade 7 FCAT were mentioned again in the Grade 9 FCAT comments. Reviewers noted that for some items it was difficult to identify the benchmarks they were targeting: item 13 (one reviewer), item 24 (one reviewer), item 26 (one reviewer), item 28 (one reviewer), and item 34 (one reviewer). The other comments related to the items being too easy for Grade 9. The items the reviewers considered too easy were 7 (two reviewers), 22 (one reviewer), 23 (two reviewers), and 29 (one reviewer). The fact that the reviewers felt these Grade 9 items were too easy is consistent with the unacceptable Depth-of-Knowledge Consistency ratings for this grade. Three out of five standards did not fully meet the Depth-of-Knowledge Consistency criterion for Grade 9.

General Comments Made by Reviewers

<u>Grade 5 Alignment Study</u>

During the debriefing discussion for Grade 5, reviewers said that they thought the alignment between the Grade 5 benchmarks and FCAT was not perfect but acceptable. Repeating what several reviewers noted when coding the test items, they said that some

of the items were difficult to assign to a specific benchmark. The reviewers suggested that changing the language of the benchmarks might help clarify their meaning and make it easier to determine the level of cognitive complexity. They also thought that even though students and teachers were becoming more familiar with the expectations described in the benchmarks, there may be a variety of interpretations of their meaning and relevance due to the overgeneralized language of the benchmarks.

When asked if they thought that the test items covered the most important topics described in the benchmarks, they said that item 34 did not really address the content of MA.D.1.2.1 and MA.D.1.2.2 but that item 39 did a better job of covering the content of those benchmarks. They also said that item 29 addressed the number of combinations but that the benchmark it targeted did not include the language describing combinations. When asked whether the levels of complexity of the items matched the levels of complexity they expected to see according to the benchmarks, they said generally yes but at the most simplistic level. They said that each benchmark should be targeted by items representing a range of levels of complexity, but a benchmark with a moderate level of cognitive complexity should be assessed by a test item with a high level of complexity.

Grade 7 Alignment Study

During the debriefing discussion concluding the Grade 7 alignment study, reviewers said that the alignment between the benchmarks and assessment at this grade level needed slight improvement and that the test items seemed to target a limited number of benchmarks. The Range-of-Knowledge Consistency rating for Grade 7 does not support this comment, however, because only Standard A did not meet the criterion fully. In terms of content they thought was underrepresented, they said that there was not as much Geometry on the test as expected.

When asked if they thought that the test items covered the most important topics described in the benchmarks, they said that, in general, the test items seemed very simplistic compared to the depth of knowledge and skill described in the benchmarks. They also commented that not all of the benchmarks were tested but that perhaps this was by design. They said that the language of the benchmarks does not reflect the complexity of the concepts embedded within them, such as similarity and parallelism. For example, benchmark MA.C.2.3.1 does not clearly indicate the need to understand the relationship of geometric concepts or to use those in problem solving. However, the test items targeted to this benchmark have students using these concepts.

The reviewers also thought that the vague and ambiguous language used in the benchmarks, such as “understands” or “uses,” does not adequately describe the knowledge and skills students are expected to master. They also said that educators have difficulty interpreting the benchmarks and determining how students will demonstrate mastery. Consequently, they have difficulty designing instruction. They said the benchmarks should be as specific, clear, and transparent as possible and use key complexity terms, such as *analyze*, *plan*, and *design*.

When asked whether the levels of complexity of the items matched the levels of complexity they expected to see according to the benchmarks, they said the complexity of the items was mainly low or moderate while several of the benchmarks were of high complexity. They noted, however, that in some cases, such as MA.C.1.3.1, the benchmark was at a lower level of complexity than the test items. They said that if the benchmarks were written at the highest level of complexity then the instruction based on those benchmarks would include the content at the lower levels of complexity.

The following are some additional reviewer comments made after the Grade 7 FCAT study:

- Because the benchmarks describe what students are supposed to master by the end of a grade grouping, in this case Grade 8, the complexity levels of the benchmarks may be of a higher level than the test items designed for seventh-graders.
- The test items they reviewed on the Grade 7 FCAT were designed around Bloom's two-tiered cognitive model, and using a three-tiered model based on cognitive complexity will raise the complexity expectations.
- Without the inclusion of performance items, it may be more difficult to design multiple choice or gridded-reponse questions that are at a high level of cognitive complexity.

Grade 9 Alignment Study

During the debriefing discussion after the Grade 9 alignment study, the reviewers said that the alignment between the benchmarks and assessment at this grade was acceptable but could be improved. They suggested that the use of the levels of cognitive complexity would be beneficial. They also said that training teachers in the levels of cognitive complexity would make them more aware of how those levels are reflected in their own assessments.

When asked if they thought that the test items covered the most important topics described in the benchmarks, they said that the Measurement and Number Sense standards were not well represented on the test and that they did not identify any items targeting MA.E.3.4.1 or MA.E.3.4.2. They said that item 34 did not seem to target any of the benchmarks and that item 29 did not match the language of MA.E.2.4.1. (The item addressed a skill that would be needed to master the benchmark but did not represent the content of the benchmark itself.)

When asked whether the levels of complexity of the items matched the levels of complexity they expected to see according to the benchmarks, they said that the Algebra items seemed to be at a low level of complexity and that many of the benchmarks were at higher levels of complexity than the items intended to test them. They also said that more complex multiple choice and gridded-response items were needed at this grade level. The group also agreed with the reviewer who noted the Source-of-Challenge problem with item 23 because the formula to solve the problem was giving in item 45.

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 benchmarks assigned to test items. The following table shows the reliability measures for the Mathematics alignment study.

Reviewer Reliability

Grade Level	Intraclass Correlation for FCAT Items	Pairwise Agreement for Standards	Pairwise Agreement for Benchmarks
Grade 5	0.8891	0.772	0.4456
Grade 7	0.8455	0.7371	0.4308
Grade 9	0.7369	0.7242	0.4765

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 Grades 5 and 7 had good correlation.

The reviewers indicated that the most difficult aspect of the alignment study was assigning the appropriate benchmark(s) for each test item. 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 test item 16. To get the pairwise agreement for the benchmarks for the whole grade-level study, the pairwise agreement for the benchmarks is averaged over all the assessment items (Webb, 2005, p. 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 low 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 low .7 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 Mathematics Alignment Study

Grade 5 Table 5.13
Grade 7 Table 7.13
Grade 9 Table 9.13

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

Appendix B

Web Alignment Tool Tables

Grade 5	**Tables 5.1-5.12**
Grade 7	**Tables 7.1-7.12**
Grade 9	**Tables 9.1-9.12**

(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.)