

SCOTT FORESMAN READING STREET

BENCHMARK ITEM VALIDATION STUDY 2005

Gatti Evaluation, Inc.

Guido G. Gatti

Principal Investigator

In Collaboration with

Research Associates from the Wisconsin Center
for Educational Research

Consulting Team

Harry S. Hsu, Anthony J. Nitko, John Smithson

0605937

**SCOTT FORESMAN READING STREET BENCHMARK
ITEM VALIDATION STUDY 2005 (SF-BIVS-R05)**

10-30-05

Principal Investigator

Guido G. Gatti
Gatti Evaluation, Inc.
162 Fairfax Rd.
Pittsburgh, PA 15221
gggatti@comcast.net

Primary Stakeholder

Funded by Pearson Scott Foresman

For Information from Primary Stakeholder, Please Contact

Marcy Baughman
Director of Academic Research
(617) 671-2652

TABLE OF CONTENTS

EXECUTIVE SUMMARY	3
I. INTRODUCTION.....	4
II. METHODOLOGY.....	5
III. RESULTS	7
Table 1. SF-BIVS-R05 Alignment Results.....	7
IV. CONCLUSIONS AND RECOMMENDATIONS	8
Recommendations	9
Caveats	9
A.1 Surveys of Enacted Curriculum Alignment Evaluation Model	10
A.2 SEC K–12 English Language Arts Taxonomy	12
A.3 Reading/Language Arts Item Quality Checklist	16
A.4 Percent of Coding Differentials Matching in at Least a Single Topic and Topic Expectation Tandem for Ten States’ English Language Arts Objectives and the Unit Test Questions	17

EXECUTIVE SUMMARY

The ultimate goal of this project was to ensure that elementary school teachers across the United States are presented with high-quality, well-aligned *Scott Foresman Reading Street* Unit Benchmark and End-of-Year Tests to reliably monitor student progress in achieving state English language arts objectives. With the No Child Left Behind Act tying federal funding to student performance on state achievement tests, K–12 content alignment is one of the most important educational issues in the United States today. The consumers of educational materials are becoming increasingly savvy, realizing that a disconnect in curriculum-to-standards alignment is a disadvantage on test day and does not help in meeting AYP demands.

The project was ambitious, attempting to collect data and evaluate the alignment between 1,879 test questions and 6,038 educational objectives across ten states. The principal investigator worked closely with consultants from the Wisconsin Center for Educational Research (WCER), the developers of a prominent alignment evaluation model endorsed by the Council of Chief State School Officers (CCSSO), the Institute for the Education Sciences (IES), and the National Science Foundation (NSF), to ensure a fair, efficient, and independent evaluation.

Test quality and alignment results were very good for the *Scott Foresman Reading Street* Unit Benchmark and End-of-Year (EOY) Tests. Ninety-eight percent of the Unit and EOY tests aligned above the median for recently aligned state assessments. In addition, the content experts saw few test question quality issues (i.e., 49/1879). In light of this positive evidence of quality and universal content coverage, the principal investigator recommends these tests for use in classrooms across the United States to inform instruction.

Please note that the principal investigator has included in the report recommendations concerning the performance level, format, and content of the test questions.

I. INTRODUCTION

Pearson Education collaborated with Gatti Evaluation and a group of renowned assessment experts¹ to conduct quality assurance and content validation research on the questions in its 2006–07 *Scott Foresman Reading Street* Unit Benchmark and End-of-Year Assessments. The ultimate goal of this effort (SF-BIVS-R05) was to ensure that elementary school teachers across the United States are presented with high-quality, well-aligned classroom assessments to reliably monitor student progress in developing priority skills² and achieving state³ reading educational objectives.

The ultimate goal of the *Scott Foresman Reading Street* Benchmark Item Validation Study was to ensure that elementary school teachers across the United States are presented with high-quality, well-aligned classroom assessments to reliably monitor student progress in developing priority skills and achieving state reading educational objectives.

Alignment is an important aspect of the validity of assessments designed to track student achievement. Alignment has been defined as “the degree to which a set of educational objectives and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do.”⁴ The concept that the course content, instruction, and assessments students are to be held accountable to should be properly aligned to clear educational objectives is as old as education itself.⁵ With the *No Child Left Behind Act* (NCLB) tying federal funding to student performance on achievement assessments, greater importance is currently being placed on K–12 alignment issues than ever before.⁶

With the *No Child Left Behind Act* tying federal funding to student performance on achievement assessments, K–12 content alignment is one of the most important educational issues in the United States today.

The increased responsibility to ensure student performance and progress calls for close scrutiny of the alignment between what is happening in the classroom with what is happening on test day. It is now necessary for curriculum and test developers to continually work to perfect the alignment between the content of their educational materials and the changing educational objectives that define achievement. The consumers of educational materials are becoming increasingly aware that any disconnect in alignment does not help in meeting AYP demands.

The Council of Chief State School Officers (CCSSO)⁷ has funded the development of alignment evaluation models because, they write, “Methods of measuring and reporting on alignment can allow all parties to see where objectives and assessment intersect and where they do not.”⁸ A handful of alignment evaluation models have been approved jointly by the CCSSO, the Institute for Education Sciences (IES), and the National Science Foundation (NSF) for use both in program evaluations and by states to meet federal requirements for alignment between assessments and standards. The principal investigator chose one of the most prominent of these models for this study and worked closely with its developers to ensure a fair, efficient, and independent evaluation of the content covered by the 2006–07 *Scott Foresman Reading Street* Unit Benchmark and End-of-Year Assessments.

1. Tse-chi Hsu, Ph.D., Research Methods Expert [Professor (retired), Research Methodology, University of Pittsburgh]; Tony Nitko, Ph.D., Classroom Assessment Expert [Professor (retired), Research Methodology, University of Pittsburgh]; John Smithson, Ph.D., Curriculum and Assessment Alignment Expert [Research Associate, WCER, University of Wisconsin-Madison].

2. *Scott Foresman Reading Street* 2007. Pearson Education, Inc.

3. AZ, CO, FL, IN, KY, NJ, NY, NC, TN, WA

4. Webb, N. L. “Alignment of science and mathematics standards and assessments in four states.” Research Monograph No. 18, National Institute for Science Education Publications, 1999.

5. Crocker, L. Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3): 5–11.

6. Baughman, M. NCLB mandates. Presentation to National Middle School Conference, 2004.

7. <http://www.ccsso.org/>

8. CCSSO, 2002. *Models for Alignment Analysis and Assistance to States*.

.....
The principal investigator worked with the developers of a prominent alignment evaluation model, endorsed by the CCSSO, IES, and NSF, to ensure a fair, efficient, and independent evaluation.
.....

II. METHODOLOGY

The SF-BIVS-R project was ambitious, attempting to collect data and evaluate the alignment between 1,879 test questions and 6,038 educational objectives (ex. Florida State Language Arts Benchmarks and Grade Level Expectations) across ten states (AZ, CO, FL, IN, KY, NJ, NY, NC, TN, WA) in an eight-month time frame (February 1, 2005 to September 30, 2005). The *Scott Foresman Reading Street* curriculum offers five Unit Benchmark Tests for grade one with 40 multiple-choice questions and one open-ended written response task. Grades two through six have six Unit Benchmark Tests with 40 multiple-choice questions, two short-answer tasks, and one open-ended written response task. Each unit is meant to correspond to the skills covered in about every two chapters of the textbook. The End-of-Year Tests have 60 multiple-choice questions, two short-answer tasks, and one open-ended written response task.

.....
The SF-BIVS-R05 project was ambitious, attempting to collect data and evaluate the alignment between 1,879 test questions and 6,038 educational objectives across ten states.
.....

The *Scott Foresman Reading Street* program is based on the priority skills model. The model ensures that students receive the right instructional emphasis at each grade level. It also ensures a more accurate alignment to state standards. With this model in mind, Beck Evaluation and Testing Associates Inc. (BETA) was contracted to write test questions appropriate for test sections entitled Comprehension, Grammar-Usage-Mechanics, High-Frequency Words (Grade 1 Units 1–5, Grade 1 EOY, Grade 2 Units 1–3), Phonics (Grade 1 Units 1–5, Grade 2 Units 1–6, Grade 3 Units 1–6, Grades 1–3 EOY), and Vocabulary (Grade 2 Units 4–6, Grades 3–6 Units 1–6, Grade 2–6 EOY). Examples of questions,

directions for administration, a more detailed description of the model, as well as a list of which language arts skills each test is designed to assess, are available from Scott Foresman.

.....
The *Scott Foresman Reading Street* program is based on the priority skills model. The model ensures that students receive the right instructional emphasis at each grade level. It also ensures a more accurate alignment to state standards.
.....

Data collection was supervised jointly by Gatti Evaluation and consultants from WCER. An adapted version of the Surveys of Enacted Curriculum (SEC) alignment evaluation model was chosen for the SF-BIVS-R05 because of its efficiency, versatility, scientific rigor, and empirical nature. For a detailed description of the SEC alignment evaluation model, see Appendix A.1. The model is efficient because it treats content as a property of test questions and educational objectives separately. This aspect of the model was immediately used as the question pool and will be reused for each state version of the program. It was only necessary to code the test questions and state educational objectives once and then compare the codes for the various combinations.

The SEC model was also attractive because its methods have been researched and utilized in practice.⁹ The principal investigator contends that the SEC model is more rigorous than other models because it forces expert raters to code questions and objectives independently without knowledge of which objectives questions are written to assess. To maximize the rigor of the methodology the principal investigator required the raters code test questions then state objectives in separate batches of work that were given weeks apart. The SEC model supports the calculation of summary alignment statistics; a single meaningful number describes the degree to which a test's content matches that of an associated set of educational objectives useful in 1) demonstrating the caliber of the test, 2) informing revisions, and 3) making comparisons with other tests.

9. Bhola, D. S.; Impara, J. C.; and Buckendahl, C. W. Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3): 22–29.

The rating group¹⁰ consisted of education professionals with expertise in elementary school-level classroom practice, language arts curriculum knowledge, test question writing experience, and a strong research background. Raters attended a three-day seminar given by Dr. John Smithson to learn the coding process as well as to become familiar with the coding language and the coding tendencies of their colleagues. Raters were encouraged to discuss specific aspects of the coding process with each other, the principal investigator, and WCER consultants. It should be noted that, although codes were discussed among the raters, there was never a forced consensus on the codes assigned and each rater always made an independent decision as to how an item should be coded. Variation in the codes was both encouraged and warranted. The SEC model is versatile in that it allows raters to propose multiple codes as well as new codes for topics that do not fit the already existing list (see Appendix A.2 for a list of codes).

.....
Education experts, trained in the coding process, made independent decisions as to the quality and content for each test question and state educational objective.
.....

In addition to coding content, the raters examined each question for grammar, clarity, relevance, clues, bias, accessibility, and graphics problems (see Appendix A.3 for the question quality checklist). Determining that a test's questions were of the highest quality was considered the first hurdle for it to pass muster with the research team. When the experts encountered a problem with a question, they noted the problem and commented on how they would correct that problem. All comments were collected and shared with the Pearson Scott Foresman editorial staff so that they could effect any necessary corrections.

Determining that a test was adequately aligned to its designated educational content objectives was considered the second hurdle. The experts noted the reading/language arts topics and performance expectations they observed for each test question and state educational

objective independently of each other in accordance with the SEC alignment model. The raw coding data was shared with Pearson Scott Foresman. These data are useful for pointing out questions that do not contribute to enhancing test content alignment. Furthermore, a question may match its objective in topic but not require the expected level of performance. As a trend, this would result in a test that focuses too much on recitation and procedural knowledge and not enough on creativity and conceptual knowledge.

Test alignment indices (AI) were prepared by the WCER staff under the supervision of Dr. Smithson. An index was calculated for each pairing of grade level/band Unit and EOY tests with the associated set of state educational objectives. The objectives for some states (CO, FL, KY, and NY) are arranged in grade bands combining the skills required across multiple grade levels. Test codes were combined across grades to create appropriate grade band tests to align to these state objectives. Since the tests were created to encompass the most vital skills required by the states, an average state content construct (ASCC) was created and aligned. This artificial construct uses the average proportion of codes recognized across a sample of states. If in fact the priority skills model underlying *Scott Foresman Reading Street* is universal in its content coverage, the assessments should be well aligned to the ASCC. The ASCC analysis excludes content band portions of state educational objectives (CO GB K-4, CO GB 5-8, FL GB K-2, FL GB 3-5, FL GB 6-8, KY GB 1-3, NY GB Elementary, NY GB Intermediate). The alignment index is explained in more detail in Appendix A.1.

.....
Test alignment indices for each test at each grade level were prepared by the WCER staff under the supervision of Dr. John Smithson. Alignment indices range from 0.0 to 1.0, providing a single meaningful number useful in demonstrating the caliber of the test and making comparisons with other tests.
.....

10. Diane Haager, Ph.D., Associate Professor, Division of Special Education, California State University, Los Angeles; Lori Olafson, Ph.D., Assistant Professor, Department of Educational Psychology, University of Nevada, Las Vegas; Steve Lehman, Ph.D., Assistant Professor, Department of Psychology, Utah State University, Logan, Gregg Schraw, Ph.D, Professor, Department of Educational Psychology, University of Nevada, Las Vegas.

III. RESULTS

Appendix A.4 shows the percent of coding differentials matching in at least a single topic and topic-expectation tandem for ten states' English language arts objectives and the unit test questions. These results give important reliability information as they indicate that the experts, though independent, consistently recognized similar content. Table 1 reports alignment indices comparing Unit Benchmark and EOY *Scott Foresman Reading Street* tests with state objectives. Shaded areas represent results for grade bands. These alignment results are strong for both the Unit Benchmark and EOY tests

relative to alignment analyses conducted by WCER comparing state educational objectives to state assessments.¹¹ The alignment data indicates that more than 98% of the Unit and EOY AI samples are above the median for the state assessment sample.

Banded state AIs seem, as a population, a little lower than those for non-banded states (N = 14, Mean = 0.26, SD = 0.03, Minimum = 0.20, Maximum = 0.31, P₂₅ = 0.25, P₅₀ = 0.26 P₇₅ = 0.28), though they are still higher than those AIs observed between state objectives and state assessments. The results for the average state content construct (ASCC) are also high in comparison to

Table 1. SF-BIVS-R05 SEC Alignment Index Results

		Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Arizona	All Units	0.39	0.37	0.34	0.37	0.40	0.41
	EOY	0.35	0.39	0.35	0.40	0.40	0.41
Colorado	All Units		0.25		0.31		
	EOY		0.27		0.29		
Florida	All Units		0.29		0.29		
	EOY		0.23		0.25		
Indiana	All Units	0.31	0.31	0.39	0.36	0.39	0.37
	EOY	0.30	0.31	0.37	0.34	0.35	0.31
Kentucky	All Units		0.24		0.23	0.29	0.27
	EOY		0.19		0.21	0.28	0.27
New Jersey	All Units	0.17	0.19	0.24	0.23	0.24	0.28
	EOY	0.19	0.19	0.23	0.19	0.22	0.25
New York	All Units		0.25		0.27		
	EOY		0.25		0.26		
North Carolina	All Units	0.22	0.25	0.30	0.27	0.29	0.28
	EOY	0.18	0.24	0.27	0.28	0.23	0.27
Tennessee	All Units	0.25	0.25	0.34	0.35	0.40	0.38
	EOY	0.22	0.23	0.36	0.33	0.33	0.34
Washington	All Units	0.26	0.24	0.33	0.33	0.38	0.36
	EOY	0.24	0.28	0.31	0.31	0.36	0.36
ASCC	All Units	0.33	0.35	0.38	0.37	0.41	0.41
	EOY	0.29	0.33	0.37	0.33	0.35	0.37
Unit with EOY		0.58	0.63	0.64	0.64	0.70	0.70
All Units	N = 47	Mean = 0.30 SD = 0.06		Minimum = 0.17 Maximum = 0.41		P ₂₅ = 0.25 P ₅₀ = 0.29 P ₇₅ = 0.36	
EOY	N = 47	Mean = 0.29 SD = 0.06		Minimum = 0.18 Maximum = 0.41		P ₂₅ = 0.23 P ₅₀ = 0.28 P ₇₅ = 0.34	
State Assessments	N=12	Mean = 0.21 SD = 0.10		Minimum = 0.11 Maximum = 0.42		P ₂₅ = 0.14 P ₅₀ = 0.18 P ₇₅ = 0.31	

This table and its contents are proprietary information belonging to Gatti Evaluation, Inc.

Note: Average state content construct (ASCC) is the average proportion of content codes recognized across the sample of states. Grade bands are omitted from the ASCC analysis. State assessments were aligned to state English language arts standards for six states in grades K–8 between 2003 and 2005.

the AIs observed between state objectives and state assessments. The ASCC AIs are a little above half the value of the AIs between the Unit and EOY tests. Additionally, the analyses from the WCER consultants found that the Unit Benchmark and EOY test questions generally have a slightly lower performance expectation than that of the state educational objectives. A majority of the test questions, predominately in the multiple-choice format, required recall level performance while the majority of the content codes for the state objectives reflected the higher demonstrate/explain performance level. The analyses also found that the benchmark questions assessed little or no writing processes or oral communication content.

.....
Ninety-eight percent of *Scott Foresman Reading Street* tests aligned above the median for recently aligned state assessments.
.....

IV. CONCLUSIONS AND RECOMMENDATIONS

The alignment to state English language arts objectives results were very favorable. The test alignment results indicate a plane of content alignment and coverage well above that previously achieved by state assessments.¹² The high average state content construct (ASCC) alignment results demonstrate that the priority skills model underlying *Scott Foresman Reading Street* is sufficiently universal in its approach to content coverage. These results, combined with the fact that experts saw few test question quality issues, are very impressive when one considers that the benchmark tests are low-stakes assessments offered with *Scott Foresman Reading Street* and intended to inform instruction.

.....
“The consistently high levels of alignment to state and grade-specific standards indicate [*Scott Foresman Reading Street*] Unit and End-of-Year Tests are largely successful in covering content emphasized by the specific state standards analyzed.”

—Dr. John Smithson, WCER
.....

11. Between 2003 and 2005, research associates at the WCER aligned twelve pairings of elementary and middle grade state reading/language arts objectives to state assessments (e.g., they aligned 2003 Grade 6 AIMS to 2003 AZ Reading & Writing Standards) for six states.
12. Gamoran, A.; Porter, A.C., Smithson, J.; and White, P.A. Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, 19(4).

RECOMMENDATIONS

Test quality and alignment results are very good for the *Scott Foresman Reading Street* Unit Benchmark and End-of-Year Tests with respect to the study sample of state educational objectives. In light of this positive evidence of universal content coverage provided by the priority skills model, the principal investigator recommends these tests for use in classrooms across the United States to inform instruction.

Since it is the contention of the principal investigator that curriculum developers should continually work to perfect the agreement between the content of their educational materials and the state educational objectives that define achievement, it is recommended that Scott Foresman utilize the data provided by Gatti Evaluation and the WCER consultants, as per this study, to continue to improve both the quality and alignment of the questions and tests as a whole. The principal investigator specifically recommends that several of the questions currently coded at the recall performance level be modified to reflect the higher demonstrate/explain performance level required by the majority of the state objectives. The principal investigator also recommends that Scott Foresman develop and add to the *Reading Street* program benchmark tests designed to cover writing processes as well as oral communication content. Seven of the ten sets of state English language arts objectives studied here have explicit sections for oral communication content and all ten have broad writing standards. These additional benchmark tests may take different and varied formats to accommodate what can be difficult content to assess.

CAVEATS

It should be noted that evaluating quality and alignment are steps in the test validation process. The benchmark tests show a high degree of question writing quality and alignment to state educational objectives. This may be sufficient evidence that the tests can be used to inform instruction of those state objectives. It is not solely sufficient, however, for making high-stakes judgments about student achievement or predicting performance on state tests.

The coding process used to collect data is subjective in that different experts may assign different content codes. The main issue with the data collection process used in this study is that the experts find and code all the content in both the test questions and educational objectives. Three is the least number of experts recommended by WCER. More expert raters would tend to increase the quantitative alignment indices since there would be a greater likelihood of matching codes.¹³ The alignment results for the three raters are positive and would be expected to increase if more raters were used.

13. Gatti, G. "The Cumulative Advantage of Additional Independent Coders on Recounting All Available Content in State Mathematics Standards." Paper presented at the American Evaluation Association Conference in Toronto, Canada. October, 2005.

Appendix A.1

Surveys of the Enacted Curriculum Alignment Evaluation Model

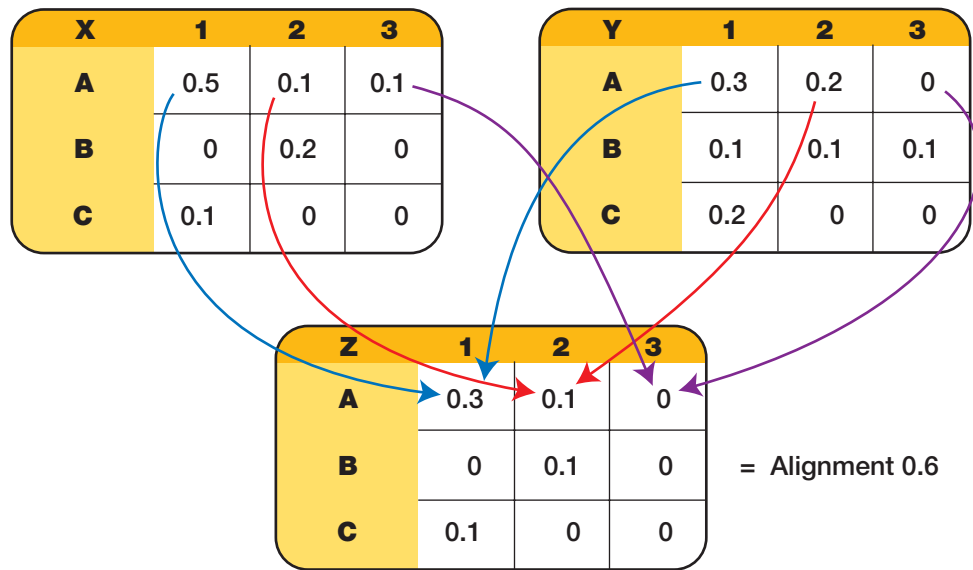
The alignment evaluation model is based upon procedures developed by Andrew Porter and John Smithson during the latter part of the 1990s. The procedure has demonstrated a strong relationship between alignment and student achievement gains¹⁴ and is one of the few approaches to alignment analyses approved by the Institute for Education Sciences (IES) for use by states in meeting federal requirements for alignment between assessments and standards. The model is also approved by the National Science Foundation (NSF) for use in program evaluations, and was developed in large part with NSF support.

The procedure utilizes a neutral, content-based taxonomy for rendering systematic and quantitative descriptions of curriculum-related documents that can be analyzed for similarities and differences. The taxonomy treats subject matter as a two-dimensional construct consisting of topics and performance expectations. The performance expectation dimension of the taxonomy utilizes five categories to describe the level of cognitive performance the typical student is expected to demonstrate for specific topics. Each performance expectation category is defined using a number of descriptors. See Appendix A.2 for the complete K–12 English language arts taxonomy. A convenient way to think about this two-dimensional construct is to consider the taxonomy as a set of descriptors for “what students should know” (topics) and “what students should be able to do” (performance expectations).

Each assessment is analyzed by at least three content experts, who use the taxonomy to write descriptions of content. While the experts are encouraged to discuss the complexities and nuances of the descriptions, each rater makes independent judgments for each element of the description. The descriptions are then combined to provide a single description of each test form. A similar process is used with the educational objectives. Once content descriptions are collected, the data is processed for quantification. The quantification process transforms expert-rater codes into proportional values. Once completed, the values across all content descriptions for any given document will add up to one. It is on these proportional values that alignment analyses are conducted.

Conceptually, the alignment index reports a proportional measure of the instructional content held in common across two content descriptions. The calculation of the alignment measure is based upon a cell-by-cell comparison made across two separate two-dimensional matrices. The figure on page 11 offers a simple example of two such matrices. Note that the values arrayed in each matrix sum to 1.0. Each matrix represents a content description. Each cell of the matrix represents a particular intersection of instructional topic by performance expectation category.

To determine the level of alignment between two such sets of data, a cell-by-cell comparison is made for each corresponding cell of the two matrices. Thus the value in cell A1 for matrix X (0.5) is compared to the value for cell A1 in matrix Y (0.3). The alignment measure reports the amount of instructional content held in common. This value is equivalent to the smaller of the two values in the comparison (in this case, 0.3). The process is repeated for each pair of cells in the matrices, with the value



held in common for each pair of cells (the smaller of the two numbers in the comparison) summed across all cells to produce the alignment measure. For the example provided in the figure, the resulting alignment value is $AI = 0.6$.

An alignment index can be calculated for any two content-based documents that have been rendered into a proportion-based record of content descriptions. Two content descriptions that are perfectly aligned will have an alignment index of 1.0. If two descriptions do not align at all, $AI = 0.0$. An index of 0.0 indicates that there is no content in common across the two descriptions. Thus, alignment indices range between 0.0 and 1.0. While there are not established criteria for what represents “good” alignment in an absolute sense, results from analyses conducted across a number of states over the past three years provide a normative basis for considering alignment values.¹⁵

14. Smithson, J.L.; and Porter, A.C. From policy to practice: the evolution of one approach to describing and using curriculum data. In M. Wilson (Ed.), *Towards Coherence Between Classroom Assessment and Accountability*. The 2004 yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

15. *ibid.*

Appendix A.2

SEC K–12 English Language Arts Taxonomy

English Language Arts Topic Areas

Strand	Content Area	Code	Topic	
Language Study	Language Study	1200	General	
		1201	Syllabication	
		1202	Spelling	
		1203	Capitalization and punctuation	
		1204	Signs and symbols (i.e., semiotics)	
		1205	Syntax and sentence structure	
		1206	Grammatical analysis	
		1207	Standard and non-standard language usage	
		1208	Linguistic knowledge (including dialects and diverse forms)	
		1209	History of language	
		1210	Relationship of language forms, contexts, and purposes (e.g., rhetoric, semantics)	
	1211	Aesthetic aspects of language		
	1290	Other		
Oral Communication	Listening and Viewing	1300	General	
		1301	Listening	
		1302	Viewing	
		1303	Nonverbal communication	
		1304	Consideration of others' ideas	
		1305	Similarities/differences of print, graphic, and nonprint communications	
		1306	Literal and connotative meanings	
		1307	Diction, tone, syntax, convention, rhetorical structure in speech	
		1308	Media-supported communication	
		1390	Other	
		1391	Listening comprehension	
		Speaking and Presenting	1400	General
			1401	Speaking and conversation
			1402	Public speaking, oral presentation
			1403	Demonstrating confidence
			1404	Effective nonverbal skills (e.g., gestures, eye contact)
			1405	Knowledge of situational and cultural norms for expression
			1406	Conversation and discussion (e.g. Socratic seminars)
			1407	Debate and structure of argument
	1408		Dramatics, creative interpretation	
	1409	Media-supported communication		
	1490	Other		
Reading	Author's Craft	800	General	
		801	Theme	
		802	Purpose (e.g., inform, perform, critique, appreciation)	
		803	Characteristics of genres	
		804	Point of view (i.e., first or third person, multiple perspectives)	
		805	Literary devices (e.g., simile, metaphor, hyperbole, flashbacks, structure, archetypes)	
		806	Literary analysis (e.g. symbolism, voice, style, tone, mood)	
		807	Influence of time and place on authors and texts	
		890	Other	

Awareness of Text and Print Features	400	General
	401	Book handling
	402	Directionality
	403	Parts of a book (e.g., cover, title, front, back)
	404	Letter and word distinctions
	405	Punctuation
	406	Text features (e.g., index, glossary, table of contents, subtitles, headings, fonts)
	407	Graphics (e.g., images, illustrations)
	490	Other
	491	Environmental print
	492	Alphabetizing
	Comprehension	600
602		Phrase
603		Sentence
604		Paragraph
605		Main idea(s), key concepts
606		Narrative elements (e.g., events, characters, setting, plot, cause and effect, structure)
607		Persuasive elements (e.g., propaganda, advertisement, emotional appeal)
608		Expository elements (e.g., description, explanation, lists, cause and effect, structure)
609		Technical elements (e.g., bullets, instruction, form, sidebars)
610		Electronic elements (e.g., hypertext links, animations)
611		Strategies (e.g. prior knowledge, prediction, inference, imagery, summarization)
612		Metacognitive process (i.e. reflecting about one's thinking process)
613		Self-correction strategies (e.g. monitoring, cueing systems, and fix-up)
690		Other
691		Sequences
692	Generating questions	
693	Building/using background knowledge	
694	Passage comprehension	
Critical Reading	700	General
	701	Fact and opinion
	702	Appeals to authority, reason, emotion
	703	Validity and significance of assertion or argument
	704	Relationships among purpose, organization, format, and meaning of text
	705	Author's assumptions
	706	Comparison of topic, theme, treatment, scope, or organization across texts
	707	Inductive/deductive approaches to comprehension
	708	Logical and faulty reasoning in text
	709	Textual evidence
	790	Other
	791	Reality vs. fantasy
Fluency	500	General
	501	Prosody (e.g., phrasing, intonation, inflection)
	502	Automaticity of words and phrases (e.g. sight and decodable words)
	503	Speed/Pace
	504	Accuracy
	590	Other
	591	Independent reading

Strand	Content Area	Code	Topic	
Reading (continued)	Phonemic Awareness	100	General	
		101	Phoneme isolation	
	102	Phoneme blending		
	103	Phoneme segmentation		
	104	Onset-rime		
	105	Sound patterns		
	106	Rhyme recognition		
	107	Phoneme deletion/substitution		
	190	Other		
	191	Identify syllables		
	Phonics	200	General	
		201	Alphabet recognition	
		202	Consonants	
		203	Consonant blends	
		204	Consonant digraphs (e.g., <i>ch, sh, th</i>)	
		205	Diphthongs (e.g., <i>oi, ou, ow, oy</i>)	
		206	R-controlled vowels (e.g., farm, torn, turn)	
		207	Patterns within words	
		208	Vowel letters (<i>a, e, i, o, u, y</i>)	
		209	Vowel phonemes (15 sounds)	
		290	Other	
		291	Sound/symbol relationships	
	292	Blending		
	Vocabulary	300	General	
		301	Compound words and contractions	
		302	Inflectional forms (e.g., <i>-s, -ed, -ing</i>)	
		303	Suffixes, prefixes, and root words	
		304	Word definitions (including new vocabulary)	
		305	Word origins	
		306	Synonyms and antonyms	
		307	Word or phrase meaning from context	
		308	Denotation and connotation	
		309	Analogies	
		390	Other	
	391	Reference word meaning, spelling, etc.		
	Writing	Writing Applications	1100	General
			1101	Narrative (e.g., stories, fiction, plays)
			1102	Poetry
			1103	Expository (e.g., report, theme)
			1104	Critical/evaluative (e.g., reviews)
			1105	Expressive (e.g., journals, reflections)
1106			Persuasive (e.g., editorial, advertisement, argumentative)	
1107			Procedural (e.g., instructions, brochure)	
1108			Technical (e.g., manual, specifications)	
1109			Real world applications of writing	
1190			Other	
Writing Components		1000	General	
		1001	Purpose, audience, context	
		1002	Main ideas	
		1003	Organization	
		1004	Word choice	
		1005	Support and elaboration	
		1006	Style, voice, technique	
		1090	Other	
		1091	Writing conventions	

Writing	900	General
Processes	901	Printing, cursive writing, penmanship
	902	Pre-writing (e.g., topic selection, brainstorming)
	903	Drafting
	904	Editing for conventions (e.g., usage, spelling, structure)
	905	Manuscript conventions (e.g., indenting, margins, citations, references, etc.)
	906	Final draft, publishing
	907	Use of technology (e.g., word processing, multimedia)
	990	Other
	991	Writer's process
	992	Revising

Performance Expectation Levels for Students in English Language Arts

I. Recall/Evaluate

Provide facts, terms, definitions, conventions

Describe

Locate literal answers in a text

Identify relevant information

Reproduce sounds or words

II. Demonstrate/Explain

Follow instructions

Give examples

Summarize

Identify purpose, main ideas, organizational patterns

Check consistency

Recognize relationship

III. Analyze/Investigate

Categorize, schematize

Distinguish fact and opinion

Make inferences, draw conclusions

Generalize

Order, group, outline, organize ideas

Gather information

Compare and contrast

Identify with another's point of view

IV. Evaluate

Determine relevance, coherence, internal consistency, logic

Test conclusions, hypotheses

Critique

Assess adequacy, appropriateness, credibility

V. Generate/Create

Integrate

Dramatize

Express ideas through writing, speaking, drawing

Create/develop connections with text, self, world

Synthesize content and ideas from several sources

Integrate with other topics and subjects

Develop reasonable alternatives

Predict probable consequences

Appendix A.3

Reading/Language Arts Item Quality Checklist

Content Quality of Item Stem, Answer Choices, and Associated Text

1. **The item stem and associated text present all of the information necessary to respond to the question.**
(e.g., it is not necessary to make certain assumptions, item is free of extraneous verbiage that distracts or confuses examinee, a single question is presented)
2. **Examinee cannot correctly respond to the question without fully comprehending the associated text or understanding necessary language concepts.**
(e.g., verbal clues are avoided, answer choices are written in a similar form and arranged in random order)
3. **All of the answer choices are plausible for multiple choice items.**
(e.g., the answer choices reflect common errors and all of the answer choices are relevant)
4. **There is only one correct answer choice for multiple-choice items.**
5. **The writing task is challenging for examinees and requires usage of relevant language skills.**

Language Quality of Item Stem, Answer Choices, and Associated Text

6. **The item stem, answer choices, and associated text are free of any errors in punctuation, capitalization, and grammar.**
7. **The reading level of the item stem, answer choices, and associated text is suitable for the children being tested.**
8. **The item stem, answer choices, and associated text are free of offensive language.**
(e.g., the language does not portray offensive stereotypes or denigrate specific populations)
9. **The language used in the item stem, answer choices, and associated text is unbiased.**
(e.g., the language does not discriminate between groups of children, either inhibiting a group from answering an item correctly or favoring a group)
10. **The associated text and/or writing task will be engaging and interesting for the children being tested.**
11. **Scoring procedures and test directions are well described, easy to understand and follow, as well as appropriate for associated items.**

If the Item Has an Associated Picture:

12. **All pictures are printed and labeled clearly.**
13. **All pictures are reasonable representations and are not misleading or offensive in any way.**
14. **All pictures are appropriate to the associated text or necessary to answer the question.**

Appendix A.4

Percent of Coding Differentials Matching in at Least a Single Topic and Topic Expectation Tandem for Ten States' English Language Arts Objectives and the Unit Test

		All Raters	
	Objectives	88.1% / 64.0%	N = 6,038
	AZ	91.0% / 69.2%	N = 852
	CO	81.3% / 66.7%	N = 123
	FL	92.4% / 69.3%	N = 721
	IN	92.7% / 69.0%	N = 422
	KY	90.1% / 58.1%	N = 587
	NJ	84.6% / 57.1%	N = 687
	NY	85.6% / 65.6%	N = 90
	NC	84.9% / 58.2%	N = 491
	TN	88.1% / 62.7%	N = 1341
	WA	85.6% / 70.3%	N = 633
	Unit Tests	94.2% / 79.1%	N = 1,501
	GR. 1	93.8% / 66.8%	N = 211
	GR. 2	97.7% / 86.4%	N = 258
	GR. 3	92.2% / 80.2%	N = 258
	GR. 4	91.9% / 76.7%	N = 258
	GR. 5	93.4% / 77.5%	N = 258
	GR. 6	96.1% / 84.5%	N = 258
		Rater 2	Rater 3
Rater 1	Objectives	64.5% / 28.2%	67.0% / 34.8%
	CO	60.2% / 39.0%	73.2% / 43.1%
	FL	67.8% / 27.7%	71.3% / 39.3%
	IN	80.3% / 30.9%	8.8% / 24.5%
	KY	69.0% / 22.7%	55.4% / 23.9%
	NJ	58.5% / 27.9%	60.4% / 23.2%
	NY	64.7% / 31.8%	52.9% / 35.3%
	NC	50.1% / 22.0%	65.0% / 32.6%
	TN	62.8% / 27.1%	69.1% / 27.5%
	WA	67.9% / 36.7%	71.6% / 39.8%
	Unit Tests	69.0% / 42.0%	67.6% / 37.1%
	GR. 1	56.4% / 30.3%	70.6% / 31.3%
	GR. 2	64.7% / 47.7%	68.6% / 46.5%
	GR. 3	67.4% / 55.4%	66.3% / 46.5%
	GR. 4	69.8% / 50.0%	55.8% / 40.3%
	GR. 5	76.7% / 38.4%	67.8% / 26.7%
	GR. 6	76.4% / 28.3%	76.7% / 30.2%

Rater 2	Objectives	Rater 3	Rater 4
	AZ	70.3% / 34.8%	
	CO	80.4% / 46.9%	63.4% / 24.2%
	FL	65.9% / 39.8%	
	IN	73.1% / 33.7%	
	KY	83.9% / 52.1%	
	NJ	66.3% / 36.6%	
	NY	62.2 %/ 28.5%	
	NC	71.1% / 41.1%	
	TN	55.8% / 21.6%	
	WA	67.3% / 28.0%	
		73.6% / 43.8%	
	Unit Tests	79.3% / 55.7%	
	GR. 1	67.3% / 46.4%	
	GR. 2	87.6% / 69.8%	
	GR. 3	79.8% / 58.9%	
	GR. 4	72.9% / 50.0%	
	GR. 5	85.3% / 52.7%	
	GR. 6	81.0% / 54.7%	
Rater 3	AZ		Rater 4
			62.0% / 32.0%