



**SCOTT FORESMAN READING STREET BENCHMARK
ITEM-VALIDATION STUDY 2006 (SF-BIVS-R06)**

PROJECT REPORT

7-15-06

Principal Investigator

Guido G. Gatti

Gatti Evaluation Inc.

162 Fairfax Rd

Pittsburgh, PA 15221

(412) 371-9832

gggatti@comcast.net

Primary Stakeholder

Funded By Scott Foresman¹
Pearson Education Inc.

For Information from Primary Stakeholder Please Contact

Marcy Baughman
Director of Academic Research
(617) 671-2652
marcy.baughman@pearsoned.com

In Collaboration with

Research Associates from the
Wisconsin Center for Educational Research (WCER)²

Consulting Team

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

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

¹ <http://www.scottforesman.com/>

² <http://www.wcer.wisc.edu/>

TABLE OF CONTENTS

1.	COVER PAGE_____	1
2.	STAKEHOLDERS_____	2
3.	TABLE OF CONTENTS_____	3
4.	EXECUTIVE SUMMARY_____	4
I.	INTRODUCTION_____	5-6
II.	METHODOLOGY_____	7-9
III.	RESULTS_____	10-12
	Table 1. SF-BIVS-R 2005/06 Alignment Index Results	11-12
IV.	CONCLUSIONS AND RECOMMENDATIONS_____	13
	Recommendations	14
	Caveats	14
A.1	Surveys of Enacted Curriculum Alignment Evaluation Model_____	15-17
A.2	SEC K-12 English Language Arts Taxonomy_____	18-21
A.3	Reading/Language Arts Item Quality Checklist_____	22
A.4	Percent of Coding Differentials Matching In At Least A Single Topic And Topic Expectation For Twenty One States' English Language Arts Objectives And The 2006 Reading Street Test Questions_____	23-24

EXECUTIVE SUMMARY

The ultimate goal of the *2006 Scott Foresman Reading Street Benchmark Item Validation Study* (BIVS-R06), conducted by *Gatti Evaluation Inc.*, was to ensure elementary school teachers across the United States are presented with high quality well aligned 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 any disconnect in curriculum-to-standards alignment is a disadvantage on test day and does not help with meeting AYP demands.

The BIVS-R06 project was ambitious, attempting to collect data and evaluate the alignment between 1,879 test questions and 10,516 educational objectives across 21 states. Beck Evaluation and Testing Associates Inc. was contracted by Scott Foresman to write the test questions. The Reading Street program is based on the *priority skills* model, developed by the program authors, and features phonemic awareness, phonics, fluency, vocabulary, reading comprehension, and writing in appropriate amounts at separate stages of development in grades kindergarten through 6th. The principal investigator worked closely with consultants from the *Wisconsin Center for Educational Research* (WCER), the developers of a prominent alignment evaluation model approved by the CCSSO, IES, and 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 tests. More than ninety percent of Unit Benchmark and End-of-Year tests' alignment-to-state-standards results were above the median observed for state assessments recently aligned in independent WCER studies. In direct comparisons, the Unit Benchmark and End-of-Year tests exceeded or matched the alignment for state assessments in all but one of eighteen cases. In addition, the content experts saw few test question quality issues (i.e., 49/1879). In light of this positive evidence for the quality and universal content coverage, the principal investigator recommends these tests for use in classrooms across the United States to inform English Language Arts instruction, specifically vocabulary, reading comprehension, critical reading, author's craft, and language study skills.

Please note the principal investigator has included in the report, recommendations concerning the performance level, format, and content of the test questions. The WCER consultants have also prepared a detailed alignment report including an interactive EXCEL file providing visual summaries of alignment results, content maps, and full content descriptions as well as very fine grained content analyses with a click of a mouse.

This summary and its content are proprietary information belonging to Gatti Evaluation Inc.

I. INTRODUCTION

Pearson Education collaborated with Gatti Evaluation and a group of renowned assessment experts³ 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 (EOY) assessments. The ultimate goal of this effort (SF-BIVS-R) was to ensure 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*⁴ and achieving state reading educational objectives.

The ultimate goal of the Scott Foresman Reading Street Benchmark Item Validation Study was to ensure elementary school teachers across the United States are presented with high quality well aligned classroom assessments to reliably monitor student progress in developing “priority reading 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” (Webb 1999)⁵. 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 (Crocker, 2003)⁶. 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 (Baughman, 2004⁷).

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 liability to ensure student performance and progress is forcing 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 savvy; realizing that any disconnect in alignment does not help in meeting AYP demands.

³ Tse-chi Hsu PhD, Research Methods Expert [Professor (retired), Research Methodology, University of Pittsburgh]

Tony Nitko PhD, Classroom Assessment Expert [Professor (retired), Research Methodology, University of Pittsburgh]

John Smithson PhD, Curriculum & Assessment Alignment Expert [Research Associate, WCER, Univ. Wisconsin-Madison]

⁴ Scott Foresman Reading Street program authors 2005. Pearson Education Inc.

⁵ Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. Research Monograph No. 18, National Institute for Science Education Publications.

⁶ Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3), p5-11.

⁷ Baughman, M. (2004). NCLB mandates. Presentation to National Middle School Conference

The Council of Chief State School Officers (CCSSO)⁸ has funded the development of alignment evaluation models because they feel, “Methods of measuring and reporting on alignment can allow all parties to see where objectives and assessment intersect and where they do not⁹.” 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 in both program evaluations and by states in meeting federal requirements for alignment between assessments and standards. The principal investigator has chosen one of the most prominent of these models for the 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.

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.

⁸ <http://www.ccsso.org/>

⁹ CCSSO, 2002. Models for Alignment Analysis and Assistance to States.

II. METHODOLOGY

The SF-BIVS-R project was ambitious, attempting to collect data and evaluate the alignment between 1,879 test questions and 10,516 state English language arts (ELA) educational objectives (ex., Florida State Language Arts Benchmarks and Grade Level Expectations) across 21 states¹⁰. The Scott Foresman Reading Street curriculum offers five Unit Benchmark tests for grade one with forty multiple choice questions and one open ended written response task. Grades two through six have six Unit Benchmark tests with forty 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 sixty 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 10,516 educational objectives across 21 states.

The Reading Street program is based on the *priority skills* model, developed by the program authors, and features phonemic awareness, phonics, fluency, vocabulary, comprehension, and writing in appropriate amounts as beginning readers progress through subsequent grades. The *priority skills* model is an attempt to provide for an elementary reading program that is accessible to all students¹¹ and that covers vital skills featured in state educational objectives. With this model in mind, Beck Evaluation and Testing Associates Inc. (BETA) was contracted to write test questions appropriate for test sections titled Comprehension, Grammar-Usage-Mechanics, High Frequency Words (i.e., Grade 1 Units 1-5, Grade 1 EOY, Grade 2 Units 1-3), Phonics (i.e., Grade 1 Unit 1-5, Grade 2 Unit 1-6, Grade 3 Unit 1-6, Grade 1-3 EOY), and Vocabulary (i.e., Grade 2 Unit 4-6, Grade 3,4,5,&6 Unit 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 attempting to assess, are available from Scott Foresman.

The Reading Street program is based on the “priority skills” model. The model was developed by the program authors in an attempt to provide a reading program that is accessible to all beginning readers and that covers vital skills featured in state educational objectives. With this model in mind, Beck Evaluation and Testing Associates Inc. was contracted to write the Unit Benchmark and End-Of-Year test questions.

Optional Reading Fluency tests, offered with each unit, were not coded. Baseline tests and Alternative Baseline tests, offered with the Reading Street program for each grade, were not coded as well. The decision to not code the Reading Fluency, Baseline, and Alternative Baseline tests was made by Scott Foresman and was strictly budgetary. Coding these additional tests and

¹⁰ 2005 State Sample - AZ, CO, FL, IN, KY, NC, NJ, NY, TN, WA; 2006 State Sample IL, LA, MA, MD, MI, MO, OH, OK, OR, PA, WV

¹¹ Child (August, 2006). *The new thinking on teaching kids to read*. Interview of G. Reid Lyon Ph.D. by Pamela Kruger.

including their content in the content descriptions for the Unit Benchmark and EOY tests as a single complete battery of tests would surely increase coverage of the *priority skills*.

Data collection was supervised jointly by Gatti Evaluation and consultants from WCER. An adapted version of the Surveys of Enacted Curriculum¹² (SEC) alignment evaluation model was chosen for the SF-BIVS-R06 because of its efficiency, versatility, scientific rigor, and empirical nature (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 exploited as the question pool 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 is also attractive because its methods have been researched and utilized in practice¹³.

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 of each other without knowledge beforehand of which objectives questions are written to assess. 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 used). Multiple performance and topic coding pairs may be listed to fully describe all relevant content covered by each test question or educational objective. The SEC coding language is dynamic, continually evolving with each project in an attempt to provide more accurate educational content descriptions. The SEC model also supports the calculation of summary alignment statistics; a single meaningful number that describes the degree 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.

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

The rating group¹⁴ 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. Three raters were used to maximize efficiency yet still produce reliable content descriptions¹⁵. Raters attended a three day seminar given by Dr. John Smithson to train in the coding process as well as 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, with raters offering up opinions, 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 when differing opinion existed among the experts.

¹² <http://seconline.wceruw.org/SECwebhome.htm>

¹³ Bholá, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), p21-29.

¹⁴ Diane Haager PhD, Associate Professor, Division of Special Education, California State University, Los Angeles.
Lori Olafson PhD, Assistant Professor, Department of Educational Psychology, University of Nevada, Las Vegas
Steve Lehman PhD, Assistant Professor, Department of Psychology, Utah State University, Logan
Gregg Schraw PhD, Professor, Department of Educational Psychology, University of Nevada, Las Vegas

¹⁵ Gatti, G. (2005). The Cumulative Advantage of Additional Independent Coders on Recounting All Available Content in State Mathematics Standards. Paper presented at the American Evaluation Association (AEA) Conference in Toronto, Canada, October, 2005.

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). The determination that a test's questions are of 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 question quality comments were collected and shared with the Scott Foresman editorial staff so that they could effect any necessary corrections.

The determination that a test was adequately aligned to a set of state educational objectives in content was considered the second hurdle. The experts noted the English language arts topics and performance expectations they observed for each test question and state educational objective independently in accordance with the SEC alignment model. The raw coding data file was shared with Scott Foresman along with an interactive EXCEL file displaying alignment index results, content maps, and full content descriptions. These data formats are useful for comparing content descriptions between tests and objectives in both topic and expectation, as well as, pointing out individual questions that do not contribute to enhancing test alignment. The interactive formats are recommended for both reviewing and comparing content descriptions for CBEMs of interest because they allow for visual summaries as well as very fine grained analyses with a click of a mouse.

Alignment results were prepared by the consulting Wisconsin Center for Educational Research (WCER) staff under the supervision of Dr. John Smithson.

Content descriptions, content maps, and test alignment indices (AI) were prepared by the WCER staff under the supervision of Dr. John Smithson. An AI was calculated for each pairing of grade level/band Unit and EOY test with the associated set of state educational objectives. The alignment index is explained in more detail in Appendix A.1. The objectives for some states (i.e., CO, FL, IL, KY, MA, NY) are arranged in grade bands combining the skills required across multiple grade levels. For these states, test codes were combined across grades to create appropriate grade band tests to align to these state objectives.

Since the Scott Foresman Reading Street tests were created to encompass the most vital skills required by all US states, a 21 state composite content description (SCCD) was created and aligned to the Unit and EOY tests. The SCCD treats all the ELA educational objectives for the 21 states currently in the BIVS-R sample as belonging to a single set of educational objectives. Aligning the SCCD to the Unit and EOY tests provides summary information on how these tests cover the ELA content included in a large sample of states' objectives. If in fact the *priority skills* model underlying the Reading Street program is universal in its content coverage, the assessments should be well aligned to the SCCD.

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 educational objectives and the Scott Foresman Unit Benchmark test questions. These results are important reliability information as they indicate the experts, though independent, consistently recognized similar content. The content experts saw very few test question quality issues (i.e., 49/1879).

Table 1 reports alignment indices (AIs) comparing Unit Benchmark and EOY Scott Foresman Reading Street tests with state English language arts (ELA) educational objectives. For a detailed description of the alignment model and the alignment index statistic see Appendix A.1. These alignment results are strong for both the Unit Benchmark and EOY tests relative to independent alignment analyses conducted by WCER comparing a limited number of state educational objectives to corresponding state assessments¹⁶. Current alignment data indicates that more than 90% of the alignment indices for the Unit and EOY sample are above the median for the state assessment sample and more than 30% of the alignment indices for the Unit and EOY sample are above the 90th percentile for the state assessment sample.

The AI results for the 21 state composite content description (SCCD) are also high in comparison to the AIs observed between state ELA objectives and state assessments. The SCCD is a composite content description including simultaneously the educational objectives for all 21 study states. Aligning the SCCD to the Unit and EOY tests gives summary information on how these tests cover the ELA content included a large sample of states objectives. This is important information since the benchmark tests are not designed to be specific to certain state educational objectives, but rather, they are designed to inform instruction on priority reading skills across all US states. All SCCD AIs for the Unit tests for the Unit and EOY tests exceed the 90th percentile for the state assessment sample.

More than ninety percent of Reading Street benchmark tests' alignment-to-state-standards results were above the median observed for recently aligned state assessments and in direct comparisons exceeded or matched the alignment for state assessments in all but one of eighteen cases.

Nine alignment index results for four states may be directly compared to those observed for both the Reading Street Unit and EOY tests. Eight of the nine AIs for the Unit tests exceeded those for the state assessments with one tie (i.e., average difference of 0.07 or 1.17 standard deviations) and seven EOY AIs exceeded those for state assessments with one tie (i.e., average difference of 0.05 or 0.83 standard deviations). In making these direct comparisons a single caveat should be noted, although the alignment model, procedures, and data analyses are identical, the pool of expert raters used to provide the content descriptions for the state assessment sample differs from that used in the SF-BIVS-R study. Most notably, four or more raters were used to construct content descriptions for the state assessment sample while three raters were used in the SF-BIVS-R study. Using four over three raters amounts to an advantage in aligning CBEMs as more raters will generally have the effect of decreasing the occurrences of missed content codes

¹⁶ Between 2003 and 2005 research associates at the Wisconsin Center Educational Research used the Sec model to independently aligned 10 pairings of grade 3 through grade 6 state reading/language arts educational objectives to corresponding state assessments (ex., align 2003 Grade 6 AIMS to 2003 AZ Reading & Writing Standards) for five states.

making for more complete content descriptions and thus higher alignment indices (see reference 15).

Additional analyses were performed by the WCER consultants that looked at overall congruence between the Unit benchmark and EOY tests with state educational objectives in both ELA topics and performance expectation separately and broken down by content area¹⁷. These analyses found that the Scott Foresman tests were exceptionally well aligned with state educational objectives in the *vocabulary, reading comprehension, critical reading, author's craft, and language study* content areas. The analyses also found that the alignment could be improved for some tests in the content areas of *phonemic awareness, phonics, writing components, and writing applications*. The tests were found to assess little or no *fluency, writing process, awareness of text and print features, or oral communication content*.

2005 States		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	0.25		0.29		0.29	
	EOY	0.23		0.27		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	
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
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	
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
2006 States		Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Illinois	All Units	0.21		0.24		0.36	
	EOY	0.22		0.24		0.36	
Louisiana	All Units	0.24	0.19	0.22	0.18	0.20	0.24
	EOY	0.23	0.21	0.25	0.20	0.19	0.22
Massachusetts	All Units	0.33		0.41		0.39	
	EOY	0.33		0.41		0.38	
Maryland	All Units	0.24	0.24	0.22	0.26	0.26	0.28
	EOY	0.22	0.24	0.23	0.26	0.26	0.28
Michigan	All Units	0.18	0.21	0.20	0.18	0.19	0.19

¹⁷ Refers to data from overall alignment tables offered in interactive EXCEL file prepared by WCER consultants, see Table 6 under *Diagnostic Use of Alignment and Content Analyses* from Smithson, J. L. (June, 2006) summary report.

	EOY	0.23	0.22	0.17	0.17	0.16	0.16
Missouri	All Units	0.17	0.19	0.21	0.22	0.25	0.24
	EOY	0.18	0.20	0.23	0.23	0.22	0.25
Ohio	All Units	0.21	0.22	0.20	0.22	0.26	0.23
	EOY	0.19	0.21	0.21	0.18	0.22	0.21
Oklahoma	All Units	0.21	0.22	0.20	0.24	0.28	0.27
	EOY	0.21	0.24	0.22	0.24	0.27	0.30
Oregon	All Units	0.26	0.22	0.24	0.26	0.31	0.27
	EOY	0.24	0.24	0.23	0.26	0.26	0.26
Pennsylvania	All Units			0.23		0.27	
	EOY			0.23		0.23	
West Virginia	All Units	0.23	0.22	0.21	0.23	0.21	0.20
	EOY	0.22	0.25	0.24	0.27	0.24	0.18
SCCD	All Units	0.27	0.29	0.31	0.30	0.34	0.34
	EOY	0.24	0.27	0.31	0.29	0.28	0.31
All Units	N = 103 Mean = 0.26 SD = 0.06		Minimum = 0.16 Maximum = 0.41		P ₂₅ = 0.22 P ₅₀ = 0.25 P ₇₅ = 0.30		
EOY	N = 103 Mean = 0.26 SD = 0.06		Minimum = 0.17 Maximum = 0.41		P ₂₅ = 0.22 P ₅₀ = 0.24 P ₇₅ = 0.28		
State Assessments	N = 10 Mean = 0.21 SD = 0.06		Minimum = 0.14 Maximum = 0.34		P ₃₀ = 0.16 P ₅₀ = 0.19 P ₇₀ = 0.23 P ₉₀ = 0.26		

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

Note: The SCCD is a composite content description including simultaneously the educational objectives for all 21 study states (AZ, CO, FL, IN, KY, NC, NJ, NY, TN, WA, IL, LA, MA, MD, MI, MO, OH, OK, OR, PA, WV). Five states' assessments were aligned to corresponding grade 3-6 state English language arts educational objectives by WCER researchers for independent projects between 2003 and 2006. Four of these five states are included in the BIVS-R study sample of 21 (n=9).

IV. CONCLUSIONS AND RECOMMENDATIONS

The alignment of Scott Foresman Reading Street Unit Benchmark and End-Of-Year tests to state English language arts objectives was very favorable. The test alignment results indicate a plane of content alignment and coverage at or above that previously achieved by state assessments. The high state composite content description (SCCD) alignment results demonstrate that the *priority skills* model underlying the Scott Foresman Reading Street program, and the benchmark tests written for the program, are sufficiently universal in their approach to ELA content coverage. This holds most true for vocabulary, reading comprehension, critical reading, author's craft, and language study skills. These results combined with the fact that experts saw few test question quality issues (i.e., 49/1879) is very impressive when one considers that the benchmark tests are low stakes assessment, offered with the Scott Foresman Reading Street program, intended to inform instruction.

Alignment results as well as content maps and full content descriptions are available in an interactive EXCEL file¹⁸ and will also be made available on the Surveys of Enacted Curriculum website (<http://seconline.wceruw.org/SECwebhome.htm>). These interactive formats are recommended for both reviewing and comparing content descriptions for CBEMs of interest because they allow for visual summaries as well as very fine grained analyses with a click of a mouse.

“The consistently high levels of alignment to state and grade-specific standards indicate (Scott Foresman Reading Street) Unit Benchmark and End-Of-Year tests are largely successful in covering content emphasized by the specific state standards analyzed.”

Dr. John Smithson, WCER (2005)

“The Scott Foresman (Unit Benchmark and End-Of-Year) assessments appear well aligned to the content standards across the twenty-one states for which alignment analyses were conducted. The assessments consistently reveal levels of alignment across grades 1-6 that match, and often exceed, alignment results for state tests to state standards based upon available data.”

Dr. John Smithson, WCER (2006)

¹⁸ Smithson, J. L. (June, 2006) Summary report on alignment analyses of Scott Foresman language arts test forms to state language arts standards in twenty-one states. An independent research project report commissioned by Pearson Education in partnership with Gatti Evaluation Inc.

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 the principal investigator recommends these tests for use in classrooms across the United States to inform instruction. The Scott Foresman tests were found to cover *vocabulary, reading comprehension, critical reading, author's craft, and language study skills* particularly well.

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 to revise both the quality and alignment of the questions and tests as a whole. To this end, it is recommended the editors work to improve coverage of *phonemic awareness, phonics, and fluency* skills. Additional content coverage may be achieved simply by including in the content descriptions the content found in the optional Reading Fluency, Baseline and Alternative Baseline tests currently offered with the Reading Street program.

Additionally it is felt that alignment with state educational objectives could also be significantly improved through the development and/or inclusion of optional tests designed to specifically cover *writing components, writing applications, and writing processes* as well as *visual/oral communication* content. The inclusion of additional tests may be warranted considering seventeen of the 21 sets of state English language arts objectives studied here have explicit *visual/oral communication* strands and all 21 contain a broad set of writing objectives. These additional tests may take different and varied formats to better 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 to those state objectives. It is not solely sufficient, however, for making high stakes judgments about student achievement or predicting performance on state achievement tests.

It should also be noted that 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 expert raters recommended by WCER. More expert raters would generally have the effect of decreasing the occurrences of missed content codes making for more complete content descriptions and thus higher alignment indices. The alignment results for the three raters are positive and would be expected to only increase if more raters were used.

Appendix A.1

Surveys of 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 1990's. The procedure has demonstrated a strong relationship between alignment and student achievement gains¹⁹ 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 for use in program evaluations, and was developed in large part with NSF support.

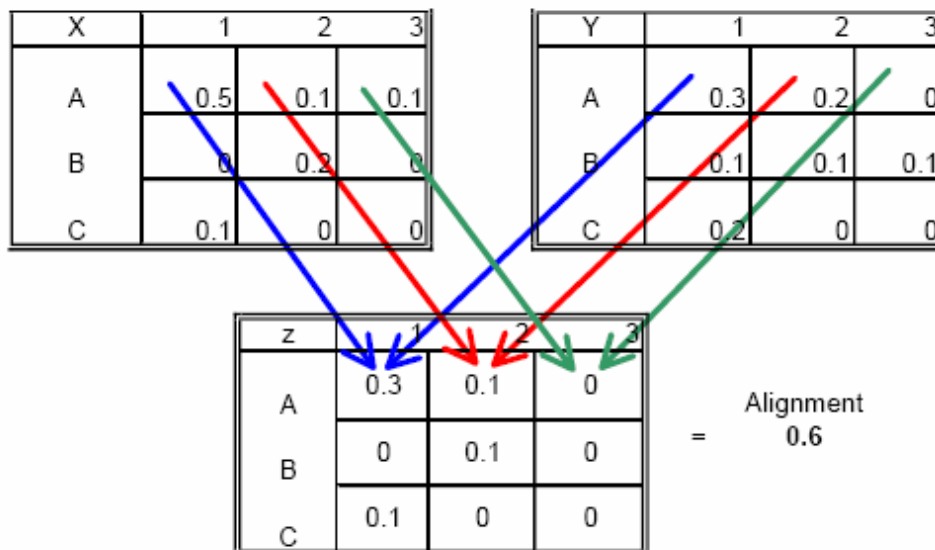
The procedure utilizes a neutral, content-based, taxonomy for rendering systematic and quantitative descriptions of curriculum-related documents that can be analyzed for similarity and differences. The taxonomy treats subject matter as a two-dimensional construct consisting of topics and performance expectations. The performance expectation dimension of the taxonomy utilizes 5 categories to describe the level of cognitive performance the typical student is expected to engage 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 'be able to do' (performance expectations).

Content based documents are rendered into descriptions of content using this taxonomy through a process of content analysis. Content analyses are conducted by content experts. Each document is analyzed by at least three content experts. While the process is collegial, and the experts are encouraged to discuss complexities and nuances of the descriptions, each rater makes independent judgments for each element of the description. The descriptions provided by the experts 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 will sum to one across all content descriptions for any given document. It is on these proportional values that alignment analyses are conducted.

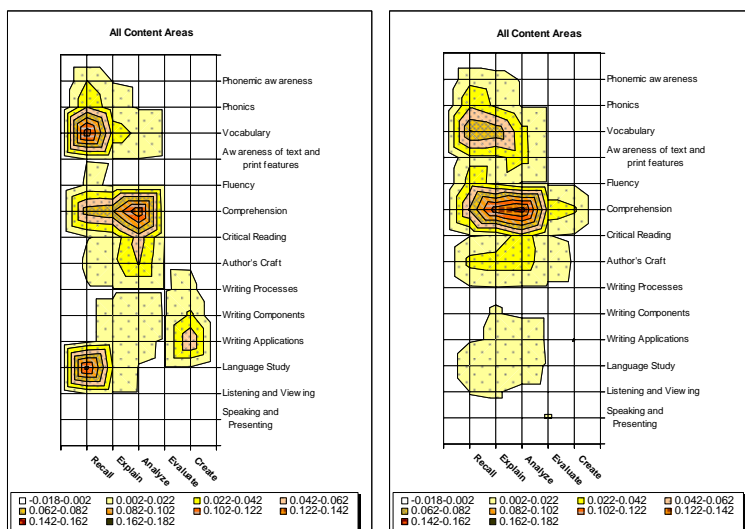
Conceptually, the alignment index reports a proportional measure of the instructional content held in common across two content descriptions. The calculation of the alignment measure is based upon a cell by cell comparisons made across two separate two-dimensional matrices. The figure below offers a 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 then repeats for each pair of cells in the matrices, with the value held in common for each pair of cells (the smaller of the two

¹⁹ Gamoran, A., Porter, A.C., Smithson, J., & White, P.A. (1997, Winter). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, 19(4).

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. (e.g., a test compared to itself). If two descriptions are perfectly mal-aligned, 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 independent alignment studies conducted across a number of states over the past several years provide a normative basis for considering alignment values²⁰.



²⁰ Smithson, J.L. & Porter, A.C. (2004). 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.

Alignment index tables, content maps (i.e., examples shown above) and full content descriptions are made available in an interactive EXCEL file as well as on the Surveys of Enacted Curriculum website (<http://seconline.wceruw.org/SECwebhome.htm>). These interactive formats are recommended for both reviewing and comparing content descriptions for CBEMs of interest because they allow for visual summaries as well as very fine grained diagnostic analyses with a click of a mouse.

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., gesture, 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	400	General	

Text and Print Features	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	General
	602	Phrase
	603	Sentence
	604	Paragraph
	605	Main idea(s), key concepts
	606	Narrative elements (e.g., events, characters, setting, plot, cause & effect, structure)
	607	Persuasive elements (e.g., propaganda, advertisement, emotional appeal)
	608	Expository elements (e.g., description, explanation, lists, cause & 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	Build/use 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	Real vs. Fantasy	
Fluency	500	General
	501	Prosody (e.g., phrasing, intonation, inflection)
	502	Automaticity of words and phrases (e.g. site and decodable words)
	503	Speed/Pace
	504	Accuracy
	590	Other
	591	Independent reading

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
201		Alphabet recognition
202		Consonants
203		Consonant blends
204		Consonant digraphs (e.g., ch, sh, th)
205		Diphthongs (e.g., oi, ou, ow, oy)
206		R-controlled vowels (e.g., farm, torn, turn)
207		Patterns within words
208		Vowel letters (a, e, i, o, u, y)
209		Vowel phonemes (15 sounds)
290		Other
291		Sounds symbol relationships
292		Blending
Vocabulary	300	General
	301	Compound words and contractions
	302	Inflectional forms (e.g., -s, -ed, -ing)
	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 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 & Associated Text

- 1. The item stem and associated text present all the information necessary to respond to 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 question without fully comprehending the associated text or understanding necessary language concepts.**
(ex., verbal clues are avoided, answer choices are written in a similar form and arranged in random order)
- 3. All the answer choices are plausible for multiple choice items.**
(e.g., the answer choices reflect common errors and all the answer choices are relevant)
- 4. There is only one correct answer choice for multiple choice items.**
- 5. Writing task is challenging for examinees and requires usage of relevant language skills.**

Language Quality of Item Stem, Answer Choices & 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 is free of offensive language.**
(i.e., language that portrays offensive stereotypes or denigrates specific populations)
- 9. The language used in the item stem, answer choices, and associated text is unbiased.**
(i.e., language that unfairly discriminates between groups of children, either inhibiting or favoring a group in answering an item correctly)
- 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 associated text or necessary to answer question.**

This checklist is a proprietary document of Gatti Evaluation Inc.

Appendix A.4

**Percent of Coding Differentials Matching In At Least A Single Topic And Topic-Expectation
For Twenty One States' English Language Arts Objectives And The 2006 Reading Street Unit
Test Questions**

		Across Three Raters ¹	
State Objectives		89.4% / 65.3%	N = 10,516
2005 Sample		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
	NC	84.9% / 58.2%	N = 491
	NJ	84.6% / 57.1%	N = 687
	NY	85.6% / 65.6%	N = 90
	TN	88.1% / 62.7%	N = 1341
	WA	85.6% / 70.3%	N = 633
2006 Sample		91.1% / 67.0%	N = 4,478
	IL	97.3% / 74.1%	N = 185
	LA	91.1% / 67.8%	N = 314
	MA	92.3% / 75.3%	N = 247
	MD	89.5% / 64.9%	N = 1,007
	MI	90.0% / 68.0%	N = 400
	MO	91.0% / 75.0%	N = 200
	OH	92.3% / 60.4%	N = 666
	OK	93.2% / 70.5%	N = 603
	OR	89.2% / 65.2%	N = 511
	PA	90.5% / 60.8%	N = 158
	WV	89.8% / 70.6%	N = 187
Unit Tests		94.2% / 79.1%	N = 1,501
	GR1	93.8% / 66.8%	N = 211
	GR2	97.7% / 86.4%	N = 258
	GR3	92.2% / 80.2%	N = 258
	GR4	91.9% / 76.7%	N = 258
	GR5	93.4% / 77.5%	N = 258
	GR6	96.1% / 84.5%	N = 258
Rater 1	Rater 2	Rater 3	
2006 Sample	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%	78.8% / 24.5%	
KY	69.0% / 22.7%	55.4% / 23.9%	
NC	50.1% / 22.0%	65.0% / 32.6%	
NJ	58.5% / 27.9%	60.4% / 23.2%	
NY	64.7% / 31.8%	52.9% / 35.3	
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%	
GR1	56.4% / 30.3%	70.6% / 31.3%	
GR2	64.7% / 47.7%	68.6% / 46.5%	
GR3	67.4% / 55.4%	66.3% / 46.5%	
GR4	69.8% / 50.0%	55.8% / 40.3%	

	GR5	76.7% / 38.4%	67.8% / 26.7%	
	GR6	76.4% / 28.3%	76.7% / 30.2%	
			Rater 3	Rater 4
Rater 2	2005 Sample		70.3% / 34.8%	
	AZ		80.4% / 46.9%	63.4% / 24.2%
	CO		65.9% / 39.8%	
	FL		73.1% / 33.7%	
	IN		83.9% / 52.1%	
	KY		66.3% / 36.6%	
	NC		55.8% / 21.6%	
	NJ		62.2% / 28.5%	
	NY		71.1% / 41.1%	
	TN		67.3% / 28.0%	
	WA		73.6% / 43.8%	
	Unit Tests		79.3% / 55.7%	
	GR1		67.3% / 46.4%	
	GR2		87.6% / 69.8%	
	GR3		79.8% / 58.9%	
	GR4		72.9% / 50.0%	
	GR5		85.3% / 52.7%	
	GR6		81.0% / 54.7%	
	2006 Sample		78.7% / 36.5%	63.4% / 29.9%
	IL		84.3% / 51.9%	73.0% / 33.5%
	LA		82.2% / 36.9%	62.4% / 31.2%
	MA		76.9% / 38.9%	72.9% / 30.8%
	MD		74.5% / 33.9%	56.0% / 25.7%
	MI		77.5% / 42.0%	63.8% / 23.3%
	MO		82.5% / 37.0%	74.5% / 46.0%
	OH		82.3% / 37.7%	58.1% / 19.4%
	OK		78.3% / 30.0%	67.4% / 40.0%
	OR		76.9% / 34.1%	65.9% / 31.9%
	PA		88.6% / 43.0%	69.0% / 29.7%
	WV		75.9% / 36.4%	65.2% / 41.2%
				Rater 4
Rater 3	2005	AZ		62.0% / 32.0%
	2006 Sample			68.5% / 27.3%
	IL			67.0% / 25.4%
	LA			67.2% / 23.6%
	MA			78.9% / 40.5%
	MD			68.1% / 26.7%
	MI			68.3% / 26.5%
	MO			80.0% / 39.5%
	OH			63.5% / 24.8%
	OK			70.4% / 24.8%
	OR			65.4% / 27.4%
	PA			64.6% / 31.0%
	WV			72.7% / 23.5%

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

Notes: ¹Although four expert raters were trained in the coding procedure only three raters were employed for any coding task. The formulation 00.0% / 00.0% indicates, first, the percent of instances at least one English language arts topic was listed by two separate raters, and second, the percent of instances at least one English language arts topic and performance expectation coding tandem was listed by two separate raters. Coding data was collected in two bursts in two separate calendar years, 2005 and 2006.