

**SCOTT FORESMAN¹ MATHEMATICS BENCHMARK ITEM-
VALIDATION STUDY (SF-BIVS-M)**

PROJECT REPORT

9-7-04

Principal Investigator

Guido G. Gatti
Gatti Evaluation Inc.

Primary Stakeholder

Funded By Pearson Education Inc.

Contact Marcy Baughman, Director of Academic Research

In Collaboration with

Wisconsin Center For Educational Research (WCER)²

Consulting Team

Marty Cohen, Terry Goodman, Harry S. Hsu, Anthony J. Nitko,
Doris Redfield, John Smithson

¹ <http://www.scottforesman.com/>

² <http://www.wcer.wisc.edu/>

INTRODUCTION

Pearson Education collaborated with Gatti Evaluation, a group of independent consultants³, and the Wisconsin Center for Educational Research (WCER)² to conduct quality assurance and content validation research on the questions in its 2004-05 Scott Foresman mathematics benchmark tests. The ultimate goal of this effort was to present elementary mathematics teachers across the United States with high quality well aligned classroom assessments to reliably monitor student progress in achieving NCTM⁴ and state⁵ mathematics educational objectives. Well aligned benchmark assessments will provide teachers with examples of high quality test questions that operationalize educational objectives. The SF-BIVS-M employed the Surveys of Enacted Curriculum (SEC)⁶ alignment evaluation model.

For the purposes of SF-BIVS-M and this report, alignment will be 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)⁷. Alignment is an important aspect of the validity of any assessment designed to track student achievement. Test developers should continually work to perfect the agreement between the content of their achievement tests and the objectives that define achievement. The Council of Chief State School Officers (CCSSO)⁸ has assisted in the development of the SEC model 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.

To evaluate alignment the SF-BIVS-M trained a group of mathematics subject matter and alignment experts¹⁰ to: 1.) Rigorously analyze test question quality and accessibility (i.e., language, distracters, graphics, bias). 2.) Determine the extent to which each test question matched its associated objectives in mathematics content matter. 3.) Determine how well each set of benchmark tests covered the content of its associated set of objectives. Expert raters assigned pairs of complex predetermined codes to objectives and test questions independently. This allows for a later comparison between the codes for each question and objective. Analysis of the coding data provides information on how to improve the quality of problematic items, which items could benefit from re-alignment, and how entire tests covered the breadth of associated standards.

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]
 Doris Redfield PhD, Curriculum & Assessment Alignment Expert [Vice President of Research, AEL]
 Marty Cohen PhD, Mathematics Curriculum & Instruction Expert [Professor, Mathematics Education, University of Pittsburgh]
 Terry Goodman PhD, Mathematics Curriculum & Instruction Expert [Professor, Mathematics Education, Central Michigan State University]

⁴ <http://www.nctm.org/standards/>

⁵ FL, IL, IN, MA, NC, NJ, NY, OH, PA, VA

⁶ <http://www.secsupport.org/about.htm>

⁷ 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.

⁸ <http://www.ccsso.org/>

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

¹⁰ Bill Hopkins PhD, Project director AP Equity Initiative, former Director of Mathematics Texas education Agency
 Charles (Andy) Reeves PhD, Lecturer Mathematics Education Univ. of South Florida, former Mathematics Supervisor Florida Dept. of Ed.
 Cathy Rahlfs Sherrill MA, Lecturer Dept. of Curriculum Studies Univ. British Columbia, former Mathematics Coordinator Humble ISD Texas
 Chad Buckendahl PhD, Director, Buros Institute for Assessment Consultation and Outreach, University of Nebraska, Lincoln.
 Gregg Schraw PhD, Professor, Department of Educational Psychology, University of Nevada, Las Vegas
 Jeff Shih PhD, Assistant Professor, Department of Curriculum & Instruction, University of Nevada, Las Vegas

METHODOLOGY

The SF-BIVS-M project was ambitious, attempting to collect data and test the alignment between 7,503 test questions and 1,137 educational objectives across 10 states⁵ and a national test based on the NAEP¹¹ mathematics framework in a four month time frame (i.e., January 19, 2004 to May 15, 2005). Many questions were written to assess more than a single objective producing 9,615 question-objective pairings for which to test alignment. The Scott Foresman mathematics curriculum offers five benchmark tests in grades three to six of roughly thirty-six questions each to correspond to the skills covered in about every two chapters of the textbook. Most of the test questions for state benchmark tests were taken from the national test or borrowed from another state test. In utilizing the efficient and versatile SEC alignment model, in which content codes are a property of questions and objectives individually, it was only necessary to code the 750 national test questions, the 1,022 unique state test questions, and the 1,137 educational objectives.

Data Collection

Data collection was supervised jointly by the WCER and Gatti Evaluation. An adapted version of the SEC model was chosen for the SF-BIVS-M because of its efficiency, versatility, established scientific rigor, and empirical nature. The model is efficient because it treats content as a property of test questions and educational objectives separately. Thus the outside expert opinions collected about the content of questions and objectives can be saved and used for revising old or creating new educational materials. This aspect of the model was immediately exploited as questions from national and state test versions that were reused on other state tests were only coded once. If codes depended on both the question and the objective the question is matched to, then each question and objective pairing would need to be coded.

The SEC model is versatile in several ways but the implication for the SF-BIVS-M is that data is collected for each test question and every objective. This allows one to examine alignment at the individual test question-objective pairing level as well as any higher level (i.e., test level, grade level, grade band level). The SEC model was also attractive because its methods have been researched and utilized in practice^{9,12}. The principal investigator contends that the SEC model is more rigorous than other models because it forces 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 first code objectives then test questions and had raters code objectives and questions in separate batches of work that were given weeks apart.

The SEC model is also useful because the methodology supports the calculations of summary alignment statistics. Being able to associate a single meaningful number with the degree the test questions match objectives is useful in demonstrating the caliber of the tests, informing revisions

¹¹ <http://nces.ed.gov/nationsreportcard/mathematics/>

¹² 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.

of tests, and for comparisons across tests. WCER provided test level alignment statistics (AI) that range from 0 to 1 with 1 indicating perfect alignment. The AI statistic provides empirical quantitative evidence as to how well the collected set of content codes for an assessment matches up to those for a set of educational objectives. This metric is explained in more detail in the *Data Analysis* section.

SEC Alignment Model¹³

The alignment analyses conducted by WCER are 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 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 procedures are also approved by the National Science Foundation for use in program evaluations, and were developed in large part with NSF support.

The alignment analyses developed by Porter and Smithson utilize a neutral, content-based, taxonomy for rendering systematic and quantitative descriptions of curriculum-related documents. Because the process is systematic and yields quantitative results, descriptions can be analyzed for similarity and differences, and the results summarized using an alignment index (AI) with a range of 0 to 1 (perfect alignment = 1).

The content taxonomy utilized here treats subject matter as a two-dimensional construct consisting of topics and performance expectations. The K-8 mathematics taxonomy lists 104 topics organized into 7 content areas. The performance expectation dimension of the taxonomy utilizes 5 categories to describe the type of cognitive engagement the typical student is expected to engage for specific topics. Each performance expectation category is defined using a number of descriptors. See Appendix A.1 for the complete K-8 mathematics 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).

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 3 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. Each of the five test forms for each state and grade are aggregated to form a single description of the state or national and grade specific test. 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

¹³ Section taken from *Summary Report on Alignment Analyses of Scott Foresman Mathematics Test Forms to Gr. 4 & Gr. 8 NAEP Benchmarks & State Mathematics Standards in Ten States*, by Dr. John Smithson, WCER, University of Wisconsin-Madison, August 27, 2004

¹⁴ 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).

content descriptions for any given document. It is on these proportional values that alignment analyses are conducted.

An alignment index can be calculated for any two documents that have been rendered into a proportion-based record of content descriptions. The resulting index provides a summary measure to describe the extent of similarity between the two descriptions. Two content descriptions that are perfectly aligned will have an alignment index (AI) of 1.00. (e.g., a test compared to itself). If two descriptions are perfectly mal-aligned, $AI = 0$. An index of 0 indicates that there is no content in common across the two descriptions. Thus, alignment indices range between 0 and 1. 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 3 years provide a normative basis for considering alignment values. Results of analyses conducted on educational objectives and assessments in grades 4, 6, and 8, across ten states during 2003 yielded AIs with a range of 0.12 to 0.40, with mean $AI = 0.27^{15}$.

Expert Raters and Coding

The Scott Foresman rating group¹⁶ consisted of three education professionals with expertise in elementary school level classroom practice, mathematics curriculum knowledge, test question writing experience, and a strong research background. Raters attended a three day seminar at WCER 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 colleges. The raters coded the grade 4 NAEP framework and several tests while on site.

During the training process the raters were encouraged to discuss the coding scheme and the codes they assigned. The raters completed the bulk of the coding off site but continued to discuss specific aspects of the coding process with each other, the principal investigator, and WCER via electronic mail. It should be noted that, although codes were discussed on and off site 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. Experts will see different content in rich and complex test questions and educational objectives. The SEC model is versatile in that it allows raters to propose multiple codes as well as new codes for mathematics topics that do not seem to fit the already existing list (see Appendix A.2 for a list of new codes used for the BIVS).

In addition to coding content the SF-BIVS-M also looked at the general quality of each question. The raters examined each question for grammar, clarity, relevance, clues, bias, accessibility, and graphics problems (see Appendix A.3 for the checklist).

¹⁵ 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.

¹⁶ Bill Hopkins PhD, Project director AP Equity Initiative, former Director of Mathematics Texas education Agency
Charles (Andy) Reeves PhD, Lecturer Mathematics Education Univ. of South Florida, former Mathematics Supervisor Florida Dept. of Ed.
Cathy Rahlfs Sherrill MA, Lecturer Dept. of Curriculum Studies Univ. British Columbia, former Mathematics Coordinator Humble ISD Texas

Data Analysis

Test question quality and alignment analyses were prepared by the principal investigator from Gatti Evaluation. Each test question was scrutinized by the expert raters¹⁶ for its quality and content. The determination that a question was of highest quality was considered the first hurdle for it to pass muster with the research team. The experts looked at the general quality of each question examining each for grammar, clarity, relevance, clues, bias, accessibility, and graphics problems (see Appendix A.3 for the checklist). 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 in a single MS EXCEL¹⁷ file and shared with the Scott Foresman editorial staff so that they could effect any necessary corrections.

The determination that a question was adequately aligned to its designated educational objective in content was considered the second hurdle for it to pass muster with the research team. The experts noted the mathematics topics and expectations of each question and objective independently in accordance with the SEC alignment model. The content codes for each test question were displayed, as seen below, in separate tables for each state, grade, benchmark test, and objective. The content codes for each educational objective were displayed in separate tables in the same manner.

TOPIC CODE * COGNITIVE DEMAND CODE * UNIQUE TEST QUESTION ID NUMBER
Crosstabulation

UNIQUE TEST QUESTION ID NUMBER		COGNITIVE DEMAND CODE			Total
			C	E	
SFIL3122	TOPIC	8	0	0	8
	CODE 313	0	3	1	4
	Total	8	3	1	12
SFIL3124	TOPIC	9	0		9
	CODE 313	0	3		3
	Total	9	3		12

a. TEST STATE = IL, GRADE LEVEL = 3, BENCHMARK TEST NUMBER = 1, STATE STANDARD ID = 7AB7

Along with these tables an EXCEL file containing the raw data was shared with Scott Foresman so that they could check the alignment of the content codes for each test question-objective pairing. The data file flagged the question-objective pairings that did not match in topic alone as well as topic and expectation. Two main question-objective content matching criteria were used, one more restrictive than the other. The least restrictive criterion defined a question-objective coding match as a single code assigned to a question by any one of the three raters was seen in the codes assigned to the associated objective by any of the three raters. The more restrictive criterion was defined as the question and objective for 2 out of the 3 raters shared a single code.

These criteria are useful for pointing out questions that are poorly aligned. A question that does not meet the least restrictive criterion had experts see no content at all in common with the

¹⁷ www.microsoft.com/office

associated objective. Also consider that if topics do not match, a question is most likely not assessing what it was intended too and should be reexamined. Furthermore, a question may match its objective in topic but not reach the expected level of performance. As a trend, this would result in a test that focuses too much on recitation and performance and not enough on conceptual knowledge. Test questions and objectives with general (i.e., 100, 200, ..., 700) or new topic codes (i.e., 190, 492, etc.) were also flagged. Very general educational objectives are difficult to write questions for and new topic codes may result in mismatches simply due to coding confusion having nothing to do with content matters. Test questions with general topic codes may be problematic. These questions may be unfocused, vague, or have little or nothing to do with mathematics achievement.

To illustrate the criteria for flagging poorly aligned questions consider the following question and objective codes for raters 1, 2, and 3.

	Rater 1	Rater 2	Rater 3
Question 1	101C	101C	101D

Objective 1	101C	102C	101C

The question and objective are considered aligned in topic in the least restrictive sense because the topic code 101 was assigned to both the question and objective. The question and objective are also aligned in topic and expectation in the least restrictive sense because the topic-expectation code 101C was assigned to both the question and objective. The question is aligned in topic in the more restrictive sense because two of the three raters assigned the 101 code to both the question and objective. The question is not aligned, however, on the more restrictive criteria in topic and expectation since only rater 1 lists 101C for both the question and the objective.

Summarizing the question quality and alignment data in this way can help identify questions in need of revision. That is, is a question poorly written or not assessing the topics or performance it was meant to? The publisher is presented with independent expert opinions as to the quality, content, and alignment of each question and state objective necessary to rewrite and realign problem questions. It is also possible to identify objectives that proved difficult for the test writers.

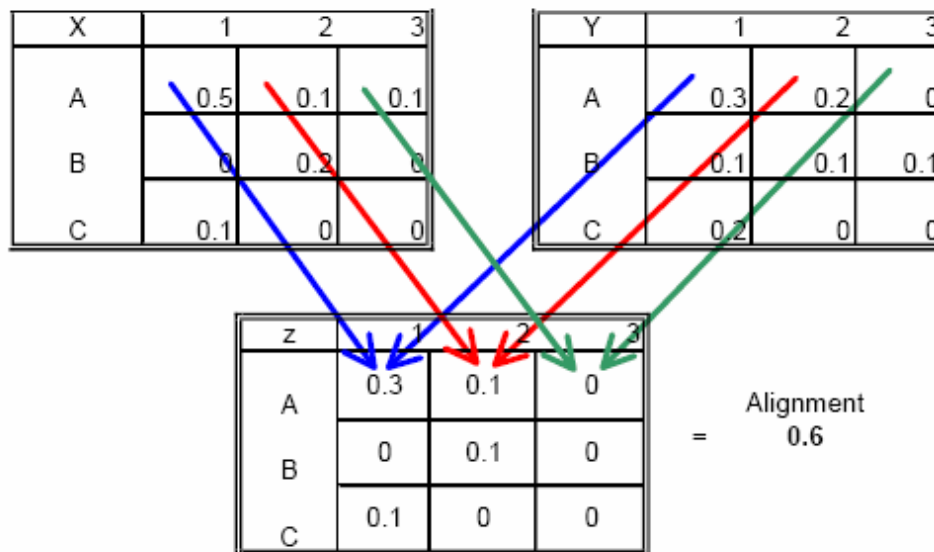
Test level alignment analyses were prepared by the WCER staff under the supervision of Dr. John Smithson³. These analyses match the content coded by trained experts for a set of educational objectives (ex., Florida grade 6 level expectations in mathematics) to that of a set of test questions (ex., end of chapter test, benchmark test, end of course test, etc.). The following description of the test level analyses is taken from a sample report written by Dr. John Smithson¹³.

The approach to measuring alignment used in this report is quantitative in nature and based upon a two-dimensional taxonomy used to describe instructional content. Conceptually, the alignment index (AI) 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. 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 (.05) is compared to the value for cell A1 in matrix Y (.03). 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, 0.03). 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 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 $AI = 0.6$.

The resulting index ranges from 0 to 1, with 0 indicating no content held in common across the two content descriptions. A measure of 1 would indicate complete agreement between the two descriptions. The measure has been used to compare instruction-to-instruction, instruction-to-assessments, assessments-to-assessments, objectives-to-assessments and objectives-to-instruction.



RESULTS

Appendix A.4 reports benchmark test to state educational objectives alignment results for each state⁵ and grade level (Note: NA = National Test based on NAEP framework). The SEC alignment index results range from 0.248 to 0.563, with a mean of 0.467 and a standard deviation of 0.07. Previously seen AI results between state assessments and educational objectives have been substantially lower¹⁵ (mean = 0.27). Across states and grades the percent of question-objective pairings that shared a common topic code ranged between 59% and 98% with a mean

of 82% and a standard deviation of 10%. The percent of pairings that shared a common topic and expectation code ranged between 56% and 95% with a mean of 76% and a standard deviation of 10%. The percent of question-objective pairings that shared a common topic code for two out of the three raters ranged between 39% and 90% with a mean of 65% and a standard deviation of 13%. The percent of question-objective pairings that shared a common topic and expectation code for two out of the three raters ranged between 28% and 71% with a mean of 47% and a standard deviation of 11%. The 95% margin of error for these percents is less than 7%.

The tests for two of the study states were written to state educational objectives that are organized by grade bands (MA, NY 3-4 & 5-6). Other state and the national tests were written to educational objectives with missing grades (PA grades 3, 5, & 8; NAEP grades 4 & 8). Though the publisher had to write some state tests to objectives written for different grades, these tests showed similar alignment. The AIs for these tests range from 0.363 to 0.533, with a mean of 0.456 and a standard deviation of 0.059. The percent of question-objective pairings that shared a common topic code ranged between 61% and 93% with a mean of 79%. The percent of pairings that shared a common topic and expectation code ranged between 50% and 83% with a mean of 69%.

The AI results for FL grade 4 and VA grade 6, though about average compared to previous analyses¹⁶, represent outliers here. The AIs are relatively low, however, a high percent of questions on these tests have content that matches their objectives. This would indicate a content coverage gap between the tests and the set of educational objectives. For example, further detailed content analyses done by WCER showed that the grade 4 FL test over-emphasizes topics associated with number sense/properties/relationships, particularly in the lower performance categories of memorize and perform procedures. These questions are easy to write for and align to basic objectives but the inclusion of too many such questions does not leave room for assessing other content put forth in the FL benchmarks. Since the AI is a measure of content coverage this over-emphasis results in a lower AI.

CONCLUSIONS

The alignment results were overall very favorable at both the question level as well as the test level. The test level statistics indicate a plane of grade level state test to state educational objectives content alignment and coverage well above that previously achieved by state assessments¹⁵. This combined with the fact that experts saw similar content, both mathematics topics and performance expectations, in 76% of the 9,615 question-objective pairings implies a great deal of attention to alignment at both the state test and question level. Scott Foresman's attention to alignment is even more impressive when one considers that the benchmark tests are low stakes assessment, offered with their textbooks, intended to inform instruction. These results also do not take into account the test question revisions made by the editorial staff that would be expected to increase both quality and alignment.

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 sufficient, however, for making 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 least number of experts recommended by WCER¹⁵. More expert raters would tend to increase the quantitative alignment indices since there is increased potential for listing more codes to match on. The alignment results for the three raters are positive and would be expected to increase if more raters were used.

Appendix A.1

K-8 MATHEMATICS TAXONOMY

Student Expectations in Mathematics

B	Memorize Facts, Definitions, Formulas
1	Recite basic mathematics facts
2	Recall mathematics terms and definitions
3	Recall formulas and computational procedures
4	_____
5	_____
6	_____
C	Perform Procedures
1	Use numbers to count, order, denote
2	Do computational procedures or algorithms
3	Follow procedures/instructions
4	Make measurements, do computations
5	Solve equations/formulas/routine word problems
6	Organize or display data
7	Read or produce graphs and tables
8	Execute geometric constructions
9	<i>Apply/Utilize Basic Knowledge/Facts</i>
10	_____
11	_____
D	Communicate Understanding of Math Concepts
1	Communicate mathematical ideas
2	Use representations to model mathematical ideas
3	Explain findings and results from data analysis strategies
4	Develop/explain relationships between concepts
5	Show or explain relationships between models, diagrams or other representations
6	_____
7	_____
8	_____
E	Conjecture, Generalize, Prove
1	Determine the truth of a mathematical pattern or proposition
2	Write formal or informal proofs
3	Recognize, generate or create patterns
4	Find a mathematical rule to generate a pattern or number sequence
5	Identify faulty arguments or misrepresentations of data
6	Reason inductively or deductively
7	<i>Spatial Reasoning</i>
8	_____
9	_____
10	_____
F	Solve Non-routine Problems / Make Connections
1	Apply and adapt a variety of appropriate strategies to solve non-routine problems
2	Apply mathematics in contexts outside of mathematics
3	Analyze data, recognize patterns
4	Synthesize content and ideas from several sources
5	_____
6	_____
7	_____

Appendix A: K-8 Mathematics Topics List

1 Number sense /Properties/ Relationships		3 Measurement		5 Geometric Concepts	
101	Place value	301	Use of measuring instruments	501	Basic terminology
102	Whole numbers	302	Theory (arbitrary, standard units, unit size)	502	Points, lines, rays, and vectors
103	Operations	303	Conversions	503	Patterns
104	Fractions	304	Metric (SI) system	504	Congruence
105	Decimals	305	Length, perimeter	505	Similarity
106	Percents	306	Area, volume	506	Triangles
107	Ratio, proportion	307	Surface Area	507	Quadrilaterals
108	Patterns	308	Direction, Location, Navigation	508	Circles
109	Real numbers	309	Angles	509	Angles
110	Exponents, scientific notation	310	Circles (e.g., pi, radius, area)	510	Polygons
111	Factors, multiples, divisibility	311	Mass (weight)	511	Polyhedra
112	Odds, evens, primes, composites	312	Time, temperature	512	Models
113	Estimation	391	Money	513	3-D relationships
114	Order of operations	392	Rate	514	Symmetry
115	Relationships between operations	393	Range	515	Transformations (e.g., flips, turns)
116	Mathematical properties (e.g., distr. property)	4 Algebraic Concepts		516	Pythagorean Theorem
2 Operations		401	Absolute value	517	Simple trigonometric ratios
201	Add, subtract whole numbers	402	Use of variables	6 Data Analysis / Prob. / Statistics	
202	Multiplication whole numbers	403	Evaluation of formulas, expressions, equations	601	Bar graph, histogram
203	Division whole numbers	404	One-step equations	602	Pie charts, circle graphs
204	Combinations of add, subtract, multiply, divide by whole numbers	405	Coordinate Plane	603	Pictographs
205	Equivalent fractions	406	Patterns	604	Line graphs
206	Add, subtract fractions	407	Multi-step equations	605	Stem and Leaf plots
207	Multiply fractions	408	Inequalities	606	Scatter plots
208	Divide fractions	409	Linear, non-linear relations	607	Box plots
209	Combinations of add, subtract, multiply, divide fractions	410	Rate of change/slope/line	608	Mean, median, mode
210	Ratio, proportion	411	Operations on polynomials	609	Line of best fit
211	Representations of fractions	412	Factoring	610	Quartiles, percentiles
212	Decimal equivalent to fraction	413	Square roots & radicals	611	Sampling, Sample spaces
213	Add, subtract decimals	414	Operations on radicals	612	Simple probability
214	Multiply decimals	415	Rational expressions	613	Compound probability
215	Divide decimals	416	Functions and relations	614	Combinations and permutations
216	Combinations of add, subtract, multiply, divide decimals	417	Quadratic equations	615	Summarize data in a table or graph
		418	Systems of equations	7 Instructional Technology	
		419	Systems of inequalities	701	Use of calculators
		420	Matrices, determinants	702	Graphing calculators
		421	Complex numbers	703	Computers and internet

Appendix A.2

NEW TOPIC CODES

- 000s problem solving as a process standard
communication as a process standard
connections as a process standard
reasoning/logic as a process standard
representations as a process standard
fact/opinion
- 190s inverses and opposites
- 290s percents, computing percentages
operations on exponents and radicals
- 390s calendar
accuracy/precision
customary measures
- 490s construct a graph or function table
multiple representations
- 590s tessellations
viewing a projection of a solid figure
map color problem
- 690s vertex-edge graphs
classification
venn diagrams
line plots
experimental design
surveys
collect data
construct hypotheses
make predictions
interpret data
- 800s problem solving

Appendix A.3

ITEM QUALITY CHECKLIST¹⁸**Content Quality of Item Stem & Answer Choices**

- 1. The item presents all the information necessary to answer the question or solve the problem clearly and concisely.**
(e.g., it is not necessary to make certain assumptions to solve problem, item is free of verbiage that distracts or confuses examinee, a single problem is presented)
- 2. The item cannot be answered without knowing the appropriate mathematical concept and/or procedure.**
(ex., verbal clues like *never* and *always* are avoided, answer choices are written in a similar form and arranged in a numerical/logical order)
- 3. All the answer choices are plausible to students that do not know the correct mathematical concept and/or procedure.**
(e.g., the answer choices reflect common student errors, all the answer choices are relevant to the problem)
- 4. There is only one mathematically correct answer choice.**

Language Quality of Item Stem & Answer Choices

- 5. The item stem is free of any errors in punctuation, capitalization, and grammar.**
- 6. The reading level of the item stem and answer choices is suitable for the students being tested.**
(The item should be free of words and concepts unknown and confusing to the average incoming student at the grade level of the benchmark test.)
- 7. The item stem is free of offensive language.**
(i.e., language that portrays offensive stereotypes or denigrates specific populations)
- 8. The language used in the item stem is unbiased.**
(i.e., language that unfairly discriminates between groups of students, either inhibiting or favoring a group in answering an item correctly)

If the item has a figure, table, or picture

- 9. The figure, picture, or table is printed and labeled clearly.**
- 10. The information presented in the figure, picture, or table is correct and/or reasonable.**
- 11. The figure, picture, or table is appropriate and necessary to answer the question.**

¹⁸ This checklist is proprietary information of Pearson Education Inc. and Gatti Evaluation Inc.

Appendix A.4

SCOTT FORESMAN BENCHMARK TEST QUESTION TO STATE EDUCATIONAL OBJECTIVE ALIGNMENT RESULTS¹⁹

		SEC Test Alignment Index	% Pairings Aligned In Topic	% Pairings Aligned In Topic By 2 Out Of 3 Raters	% Pairings Aligned In Topic And Expectation	% Pairings Aligned In Topic And Expectation By 2 Out Of 3 Raters
FL	G3	.535	93.40	77.00	83.10	60.60
	G4	.248	89.20	79.30	88.30	61.50
	G5	.498	83.50	62.90	74.70	45.10
	G6	NA	NA	NA	NA	NA
IL	G3	.527	92.00	83.50	87.10	51.40
	G4	.457	93.00	80.60	88.60	57.70
	G5	.553	82.10	65.00	75.80	50.40
	G6	.497	82.70	62.20	79.10	41.80
IN	G3	.551	81.50	69.40	79.70	48.20
	G4	.553	78.80	68.40	76.20	54.10
	G5	.563	83.10	73.80	80.00	60.50
	G6	.505	80.00	64.00	76.90	45.80
MA	G3	.509	91.20	69.40	77.80	46.30
	G4	.533	93.30	69.60	82.50	55.00
	G5	.500	85.30	67.60	78.90	55.40
	G6	.481	84.00	59.60	79.30	49.30
NA	G3	.497	84.10	67.80	66.70	43.00
	G4	.505	78.90	70.60	65.20	46.20
	G5	.387	60.80	39.90	49.70	22.60
	G6	.381	65.40	43.40	55.70	21.10
NC	G3	.468	71.30	65.50	64.50	40.90
	G4	.434	88.90	70.60	70.40	46.80
	G5	.379	82.70	58.60	63.60	36.40
	G6	NA	NA	NA	NA	NA
NJ	G3	.468	58.70	42.00	55.60	31.30
	G4	.438	61.70	51.70	58.30	38.60
	G5	.498	64.90	47.70	60.00	34.00
	G6	.449	65.00	50.20	61.10	35.60
NY	G3	.458	81.90	58.60	72.60	36.70
	G4	.487	86.40	58.40	77.40	40.70
	G5	.427	73.80	38.50	65.10	27.70
	G6	.401	75.20	40.00	65.70	29.00
OH	G3	.545	91.10	70.90	88.70	44.60
	G4	.405	82.80	69.60	75.00	52.50
	G5	.479	84.90	69.40	81.70	50.50
	G6	.485	93.20	67.50	81.90	48.60
PA	G3	.496	79.80	51.90	67.90	34.20
	G4	.363	80.00	60.40	73.80	44.90
	G5	.502	73.60	53.70	67.10	43.10
	G6	.364	65.30	46.80	62.00	31.90
VA	G3	.510	88.70	78.00	82.70	62.50
	G4	.448	94.80	79.90	87.40	69.00
	G5	.393	97.20	83.10	89.80	66.10
	G6	.304	97.70	90.20	95.40	70.70

¹⁹ NA = National Test based on NAEP framework