Describing the Academic Content of PISA Mathematics and Science Item Pools

By John Smithson, Ph.D.
Director, Surveys of Enacted Curriculum,
Wisconsin Center for Education Research, University of Wisconsin-Madison

May 2009

Send all correspondence to:
John Smithson, Ph.D.
Wisconsin Center for Education Research
University of Wisconsin-Madison

INTRODUCTION

This paper addresses two key elements in attempting to understand the nature of the Program for International Student Assessment (PISA) assessments and the message to be drawn from U.S. students' performance on these assessments relative to student performance in other nations.

Some argue that U.S. performance on PISA highlights important insufficiencies in our curriculum, suggesting that we are not covering the right content, not covering it to sufficient depth, or not presenting that content with sufficient quality to sufficiently support learning evidenced by "world-class" levels of achievement. Others counter that the PISA assessments are inappropriate measures of U.S. student progress, either because they assess the wrong content, or the right content, but using an inappropriate format or measurement strategy to assess student understanding of that content.

How are we to judge these competing claims? Most arguments hinge on the author's own analysis and interpretation and appropriateness of either the content covered or the assessment strategies employed. Differing conclusions can generally be attributed to differences in the analysis and interpretation of the salient features of these assessment instruments. Yet as consumers of these reports and editorials, we are left without access to the source documents on which the various, and often competing, assertions are based. We are left to judge the claims largely on our own predispositions relative to the issues involved or the utility of international assessments and/or rankings, and/or the rhetorical skills of the authors of these claims.

The first objective of this paper is to present descriptive features of the subject-matter content embedded in the PISA mathematics and science assessment item pools using a systematic and objective methodology developed by Andrew C. Porter (Porter & Smithson, 2001), and since adopted as a central component of the Surveys of Enacted Curriculum (SEC). The SEC online data collection, processing, and reporting system was developed and continues to be housed and maintained at the University of Wisconsin-Madison, in partnership with the Council of Chief State School Officers (CCSSO), and with initial funding from the National Science Foundation and a consortium of state education agencies. The SEC is currently in use in more than 20 states for a variety of purposes, most notably professional development, school improvement, program evaluation, and as a data source supporting basic research connecting policy, theory, and practice. One component of the SEC data system is a methodology referred to as content analysis, which employs a two-dimensional detailed content taxonomy to collect systematic and quantifiable descriptions of academic content for four subject areas: mathematics, science, English language arts and reading, and social studies (i.e., history, economics, civics, and geography). This approach to content analysis is herein referred to as the SEC methodology.

While it is not possible to present the source materials for the PISA instruments, or any other secure assessment, the SEC methodology described below provides a systematic and relatively detailed description of those source materials in a manner that communicates the curricular content embedded in the assessment items used in the PISA instruments to answer the question, just what academic content does the PISA assess?

Knowing what <u>is</u> assessed, however, is only one part of adjudicating claims concerning what <u>should</u> be assessed. Other key elements include some “target” of instruction (e.g., as described in state/national content standards, and/or measured by state/national assessments of student understanding and academic progress), and a means for determining how <u>what</u> is assessed on the PISA compares or relates to those “targets” representing what we want students to know and be able to do. This relationship between assessments and standards, and/or assessments and assessments, is sometimes referred to as alignment.

Because the SEC methodology provides a systematic and detailed language for describing academic content for any curriculum-related document (whether content standards, assessments, or instructional materials), it is a relatively straightforward procedure to compare two content descriptions in terms of their similarities and differences in content coverage. While such comparisons can be done to some extent using simple descriptive displays of PISA content compared with some relevant “target,” it is also possible to use a mathematical algorithm to make a detailed quantitative comparison of any two descriptions, and then summarize that comparison as a number on a scale running from 0 to 1, with 1 indicating a perfect match on content emphasis. This summary measure is typically referred to as the alignment index (see Comparing PISA to Standards and Other Assessments below). The second objective of this paper is to report on the relationship between PISA content and the content represented by key standards and assessments employed widely in the United States.

THE SEC CONTENT TAXONOMIES

At the heart of the Surveys of Enacted Curriculum is a two-dimensional content taxonomy designed to cover K–12 academic content in the areas of mathematics, science, English language arts and reading, and social studies. These descriptive taxonomies or content languages are based on the pioneering work of Andrew C. Porter to collect detailed and systematic descriptions of classroom content coverage on a large scale, to open up the “black-box” of classroom practice to inform research on policy effects (Porter & Smithson, 2001; Smithson & Porter, 2004).

The two dimensions of the SEC content taxonomies correspond to the familiar credo: "what students should know (topics), and be able to do (performance expectations)." The topic dimension for each subject area in SEC consists of an extensive list of subject-matter topics relevant to describing the K–12 curriculum. These topic lists are upgraded every 2 to 3 years to add new topics identified based on feedback from content experts collected over the course of numerous content analysis workshops conducted several times a year. Topics are organized into some number of content areas (ranging from 16 to 29 content areas across the 4 academic subjects). These content areas provide a means of aggregating detailed data collected at the topic level into a coarse-grain summary measure, thereby providing a means to report content emphasis at both a coarse-grain (content area) and fine-grain (topic) level.

The second dimension of the content taxonomy, "expectations for student performance," is represented by five categories of cognitive skill. Roughly speaking, these categories can be likened to Bloom's taxonomy, or Norman Webb's depth of knowledge construct. Each of these constructs is distinct, and not easily subsumed, but seek in a similar fashion some means to describe and measure the demonstration of mastery for some set of cognitive skills. The SEC content taxonomies each employ five categories of expectation of student performance (sometimes referred to as cognitive demand). Although the particular descriptors and category identifiers or labels differ from subject to subject, the five categories are intended to represent a common set of cognitive skills, which could loosely be defined as (1) recall, (2) procedural skills, (3) demonstrative understanding/application, (4) analysis, and (5) evaluation/synthesis/integration.

This two-dimensional taxonomy can be thought of as a two-dimensional matrix, with the rows of the matrix indicating topics, and the columns categories of performance expectation.

Using the SEC content taxonomies, every content matrix then has five columns and some varying number of rows (topics), depending on specific subject area and grain size being reported.

THE SEC CONTENT ANALYSIS METHODOLOGY

Although initially designed to collect information from teachers about their coverage of instructional content, Porter and his colleagues soon recognized that the same language could be used to describe assessment items, and that doing so across all the items of an assessment would yield a content description of that assessment. This technique was first employed in the *Upgrading Mathematics Study* (Gamoran et al., 1997), and was used there to demonstrate the impact of opportunity to learn on student achievement gains.

While not completely convinced that state content standards documents could be similarly described using these content languages, responding to interest among several states, and noting the increasing attention being paid to alignment of assessments to academic content standards, Porter and his colleagues first used the content taxonomies to describe state standards as part of the *Data on the Enacted Curriculum (DEC) Study*. Despite some additional and more formalized training components and exercises, the process is essentially the same as that employed initially for the DEC study. This process consists of convening a team of three to five content experts to conduct the document analysis. Preferably, content experts will have a degree in the discipline of interest, and/or substantial experience teaching the subject at the primary, secondary, or post-secondary level. We also like to include a mix of perspectives, using a combination of curriculum and assessment specialists, SEA staff, master teachers, and university faculty in the makeup of content analysis teams.

Analysts are trained in the use of the content taxonomies to describe assessment items and standard strands, and then led through a process consisting of independent, individual analysis of the document followed by a group discussion period to review items flagged by analysts during the independent coding process. Discussion is intended to provide team members the benefits of each other's rationale for selecting particular descriptions for specific items. Consensus is not required, although analysts are free to change their descriptions based on another analyst's rationale. While inter-rater reliability can be calculated, because analysts have the opportunity to share their decisions and rationale in describing the content of a document, and often do alter their own descriptions as a result, the resulting measure might be inflated because of this cross-rater sharing. Moreover, because of the flexibility permitted analysts in determining the number of descriptions to use for a given unit of analysis, determining an algorithm for measuring inter-rater reliability is not straightforward. For our purposes, we use the same algorithm to calculate alignment between two descriptions as a measure of inter-rater reliability.

Results of these content analyses are collected from each analyst and processed to provide an analyst-by-analyst description of the content embedded in the selected document. These resulting descriptions are then merged, using the mean measure across raters for each cell in the fine grain content matrix holding the content analysis data results, to generate the reported content description of that document.

DISPLAYING RESULTS

The descriptive results of these content analyses are best displayed in graphic format to present the relative content emphasis of the target document across the typological terrain represented by the content taxonomy employed. Two formats are most often used for this purpose, tile charts and topographic content maps. Examples of both are presented in Figure 1, where both displays present results from the same data table, or content matrix. While both display formats present the same information, different audiences tend to prefer one or the other format for their particular descriptive needs. Some criticize use of the surface area chart (topographic map) format with the nominal, noncontinuous variables (i.e., the topic lists and categories of performance expectation) used as the data points underlying the three dimensional topographic representations. Nonetheless, the data presented by the content maps provide slightly more visual information than available with the tile chart display, while accurately reporting the relative emphasis represented by the data points in the relevant content matrix.

PISA Descriptive Results

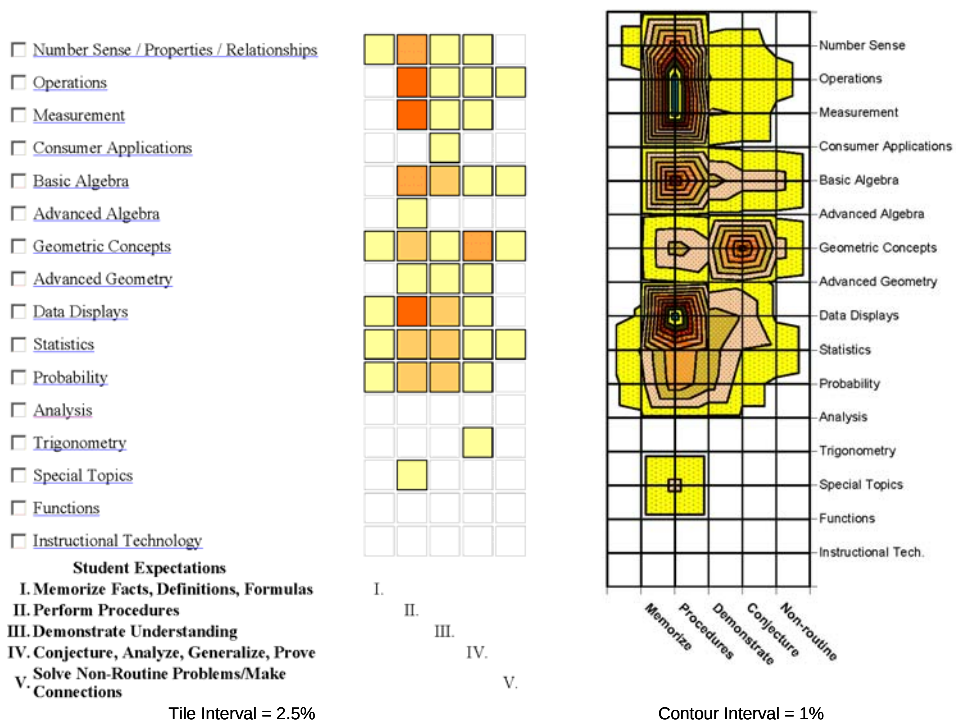
Figure 1 reports results of content analysis conducted on the complete 2006 PISA mathematics literacy item pool, displayed in both tile chart and content map formats. Figure 2 similarly reports results for the complete 2006 PISA science literacy item pool.

Results of content analyses of the 2006 PISA mathematics item pool indicate that the PISA assessments span topics across many of the content areas, and across all of the categories of performance expectations. Despite this range of assessed content represented in the mathematics PISA, 60 percent of the items emphasize procedural knowledge and skills. Nonetheless, PISA items also reveal a noticeable amount of focus at other levels of performance expectation, with the most demanding level of performance expectation (integration, synthesis, making connections) reported for four distinct content areas: Operations, Basic Algebra, Geometric Concepts, and Data Displays.

Ten of the 13 content areas represented on the PISA mathematics literacy assessments also emphasize performance expectations that demonstrate understanding of key mathematical concepts, and/or expect students to conjecture, hypothesize, or predict outcomes based on available information.

Mathematics

Figure 1



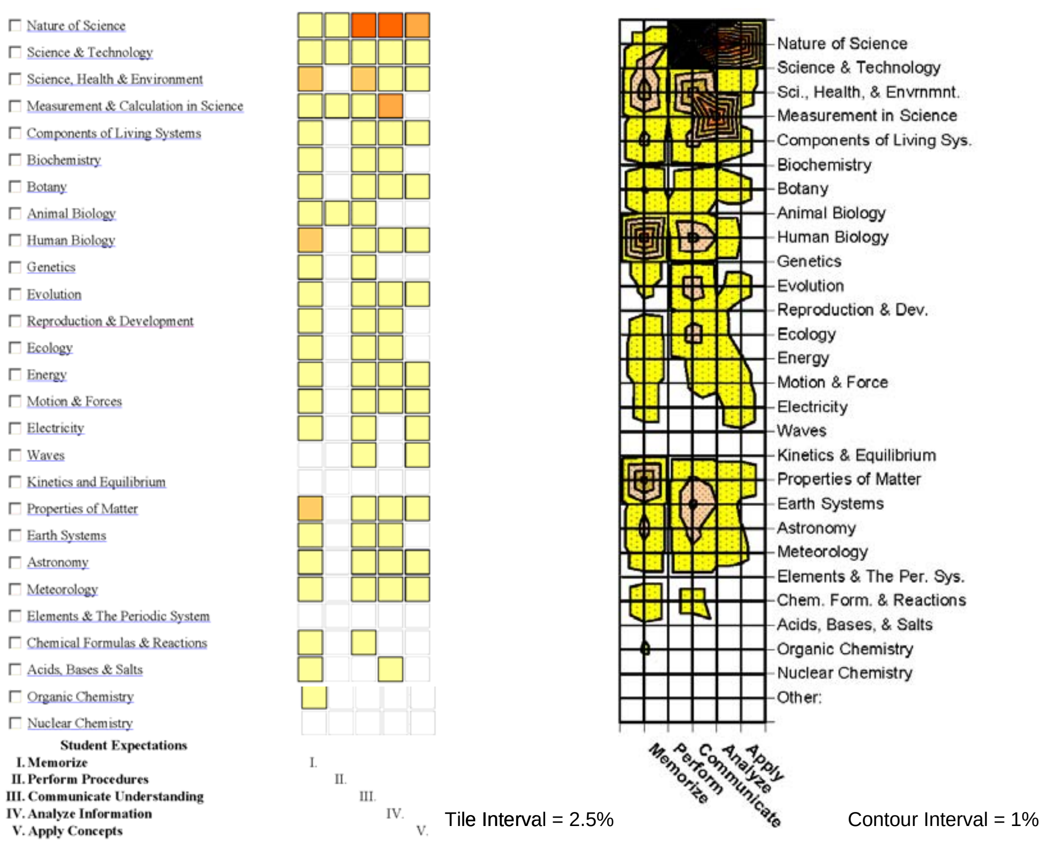
Nearly 20 percent of the PISA mathematics items focus on data displays, while an additional 50 percent of the items cover content under Operations, Measurement, Basic Algebra, and Geometric Concepts. The remaining 30 percent of content coverage is spread across Number Sense, Statistics, and Probability, with slight content coverage (less than 5 percent) associated with Advanced Geometry and Special Topics. Content covered under Special Topics refers primarily to content associated with permutations and combinations, although some items on networks are also referenced under this content area.

Science

Figure 2 reveals a broad spectrum of science content represented on the 2006 PISA science literacy item pool. Only 3 of the 27 content areas identified in the SEC science taxonomy do not have content assessed by the PISA item pool (Kinetics & Equilibrium, Elements & the Periodic Table, and Nuclear Chemistry). By far the most highly emphasized content area represented by the 2006 PISA science literacy item pool is Nature of Science, accounting for nearly one third (30 percent) of the science content represented in the item pool. Other content areas with moderate emphasis in the science item pool include Measurement in Science (8 percent), Human Biology (9 percent), and Properties of Matter (7 percent).

The heavy emphasis on Nature of Science content is combined with relatively high performance expectations, with items aimed at assessing students' ability to communicate understanding (37 percent), analyze information (22 percent), or make connections or apply to real-world situations (13 percent). About a quarter (26 percent) of the assessed content measures student recall of science concepts. Interestingly, almost no procedural knowledge skills (2 percent) are represented in the PISA science literacy item pool.

Figure 2



Although the SEC content taxonomies provide a reasonably good and detailed description of the assessed content represented in the 2006 PISA items pools, these descriptions are of course limited to the descriptors available to the content taxonomy employed. One important and well-known feature of PISA is its focus on literacy. Thus, many of the items have a heavy reading load and require some level of reading comprehension skills not directly notable within the descriptive language provided by the SEC mathematics and science taxonomies. Also, some PISA items are open-ended, requiring written responses that presuppose certain writing, organizational, and argumentative skills in addition to the student's understanding of key mathematics or science concepts involved.

While these literacy skills are not directly represented in the SEC mathematics and science content descriptions, they tend to be associated with the emphasis on more challenging performance expectations, including communicate understanding, analyze information, and make connections or solve nonroutine problems. Whether this focus on literacy skills is appropriate or not is beyond the scope of this paper or the SEC reporting system. However, the central issue in this respect may be whether the emphasis on the higher-level performance expectations can reasonably be expected to be assessed without some dependence on literacy skills. This of course presupposes that it is the goal of U.S. education efforts to prepare 15-year-olds to perform at these high-performance expectation levels and/or to possess the literacy and writing skills necessary to perform well on the PISA assessments.

Comparing PISA to Standards and Other Assessments

Because the SEC methodology for conducting content and alignment analyses renders descriptions of academic content into a common and systematic language, it is possible to do a detailed quantitative comparison of the similarities and differences between any two sets of content descriptions. Results of such analyses can be summarized using a single measure with a range of 0 to 1, and is referred to as the alignment index. A more detailed examination of alignment is possible using a set of alignment measures designed to provide detailed information about the nature of the similarities and differences in the content described by various documents, thereby serving a diagnostic function detailing the changes in one or the other content description that would lead toward higher alignment. Using the results of such alignment analyses allows one to examine the relative similarities and differences between PISA and other key educational targets at greater levels of detail than otherwise possible.

For the purposes of this paper we consider three key targets for each subject (mathematics and science) against which to compare the PISA assessment items. For mathematics these targets are the Grade 8 Trends in International Mathematics and Science Study (TIMSS) item pool, the Grade 8 National Assessment of Educational Progress (NAEP) item pool, and the National Council of Teachers of Mathematics (NCTM) Standards. For science, the comparison targets are the Grade 8 TIMSS item pool, the Grade 8 NAEP Framework (at the time of this writing the NAEP science item pool had not yet been analyzed), and the National Science Education Standards (NSES). Grade 8 targets were selected for comparison with PISA (administered to 15-year-olds), because they capture the bulk of a 15-year-old's opportunity to learn academic content, while Grade 10 and high school exit exams would likely include content not yet covered sufficiently by 15-year-olds in the United States. Thus, eighth-grade assessments and standards seem to provide a more appropriate representation of the type of academic content knowledge expected of a 15-year-old in the United States. Table 1 reports the alignment index for each combination of targets for mathematics, while Table 2 does the same for science.

Table 1

Mathematics	PISA Math	Gr.8 NCTM	Gr.8 TIMSS	Gr.8 NAEP
PISA Math	--	0.15	0.38	0.30
Gr.6-8 NCTM	0.15	--	0.21	0.22
Gr.8 TIMSS	0.38	0.21	--	0.40
Gr.8 NAEP	0.30	0.22	0.40	--

Because of the detailed nature of the content descriptions collected using the SEC procedures, the resulting alignment index becomes a very sensitive measure, with documents only rarely achieving an overall alignment index greater than 0.40. Typically, alignment measures greater than 0.30 indicate fairly good alignment. By looking at alignment numbers across a variety of different documents, however, it becomes relatively straightforward to determine which documents are more or less aligned to one another. Thus in Table 1, reporting alignment measures for mathematics across the selected documents/targets, it is clear that the assessments, whether NAEP, TIMSS, or PISA, are more similar to one another than to the NCTM standards. Comparing the three assessments with NCTM standards, the PISA assessment is the least aligned assessment to NCTM, while NAEP, as one would expect, has the highest degree of alignment to the NCTM standards. Surprisingly though, the TIMSS assessment is almost as well aligned to the NCTM standards as the NAEP assessment. In each case though, alignment of the assessment to the NCTM standards is at best modest, even for NAEP (AI = .22).

Table 2

Science	PISA Scienc	Gr.8 NSE	Gr.8 TIMSS	Gr.8 NAEP
PISA Science	--	0.35	0.38	0.33
Gr.5-8 NSE	0.35	--	0.28	0.28
Gr.8 TIMSS	0.38	0.28	--	0.45
Gr.8 NAEP	0.33	0.28	0.45	--

For science (reported in Table 2), we again see a tendency for the three assessments to be more similar to one another than they are to the national standards, although surprisingly the PISA item pool presents an exception to this pattern, because it is better aligned to the national science education standards than it is to the NAEP assessment items. Indeed, of these three assessments, PISA is better aligned to the national science standards than either NAEP or TIMSS. As noted with mathematics, the NAEP and TIMSS assessments appear better aligned to one another than either the national standards or PISA.

While tables of alignment indices can serve to identify the relative degree of similarity or difference among various curricular targets, the descriptive displays of tile charts and/or content maps are more informative in providing a sense of the nature of the differences in academic content revealed by one or another content description.

Figure 3

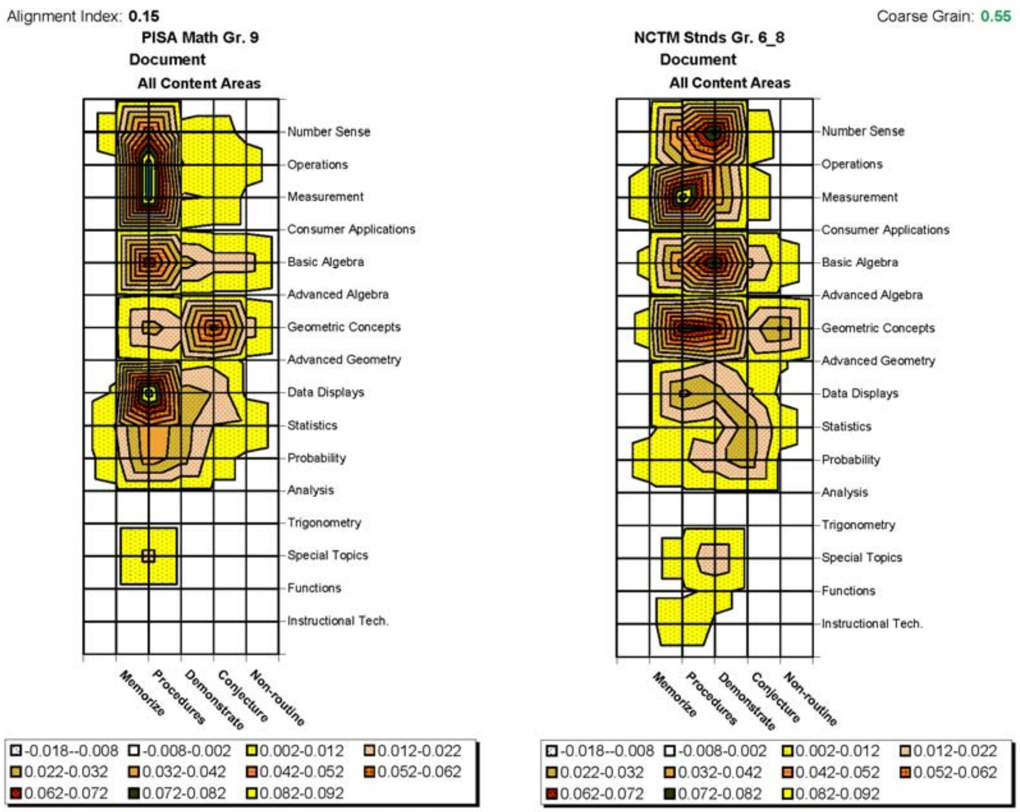
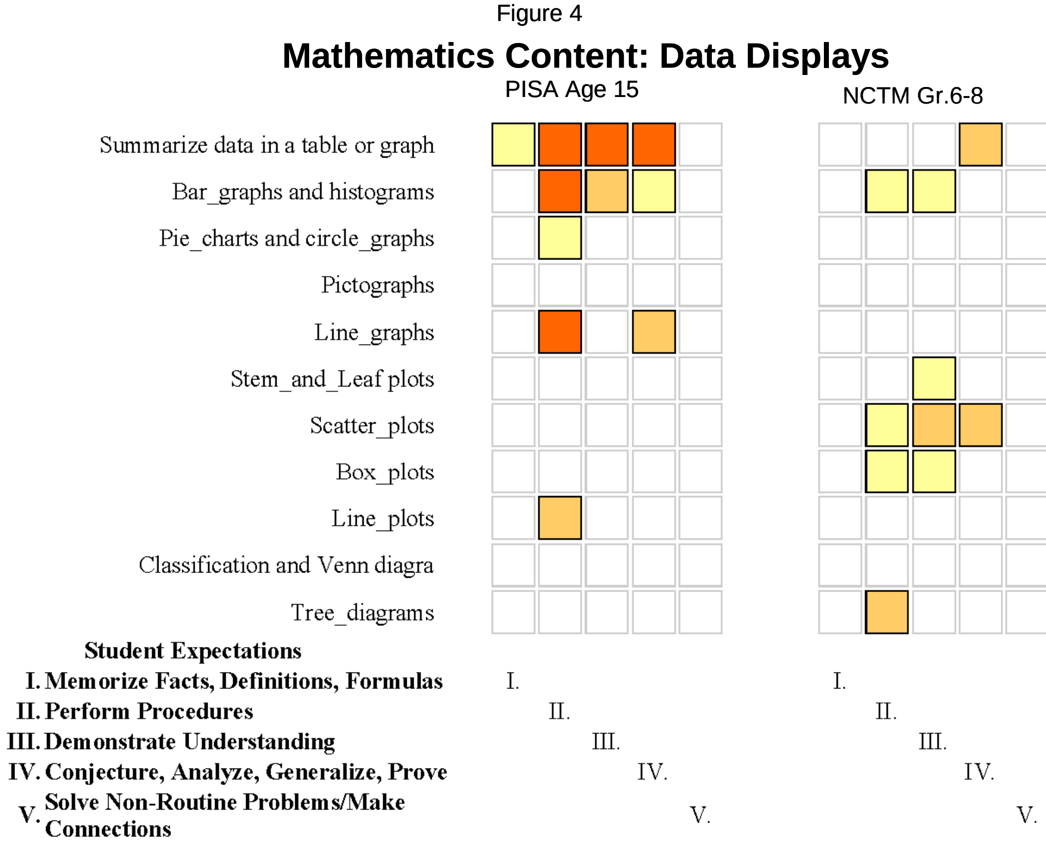


Figure 3 reports descriptive results for the PISA mathematics item pool and the content referenced in the NCTM mathematics standards. While the coarse grain maps displayed in Figure 3 do reveal some difference in emphasis on performance expectations, by and large both documents seem to cover much the same content, which seems to belie the relatively low alignment index (0.15). This apparent disparity between the alignment measure and the content overlap depicted in Figure 3 can be explained by noting that the alignment index is calculated at the fine grain level. That is to say that while there is a good deal of overlap in “content areas” as depicted in Figure 3, a more detailed look at the particular topics within those content areas would reveal that the PISA assessment and the NCTM standards target differing sets of topics within those common content areas. Note that if the alignment index is calculated at the coarse grain level (i.e., ignoring topical differences within content areas) the resulting coarse grain alignment is substantially higher (0.55).

Figure 4



An example of a mismatch at the topic level not discernible in the coarse grain map is presented in Figure 4. This tile chart reports on content related to use of data displays. As can be seen, overall PISA places a good deal more emphasis on summarizing data in a table or graph, and indeed, on data displays in general than represented in the NCTM Content Standards. Moreover, at the topic level is it quite clear that very different types of data displays are emphasized in the two content descriptions.

Should this mismatch of topics, and the low alignment index between PISA and the Grade 6–8 NCTM standards indicate that the PISA mathematics assessment is a poor measure of U.S. students' mathematical knowledge? Basing such a decision on the NCTM standards alone might be premature. A state would be well advised to compare a content map of its state academic content standards with the PISA content analysis results. Similar results for state content standards are available online for about 20 states, and are publicly accessible through the SEC Online website (http://seconline.wceruw.org/secStandards.asp).

Figure 5

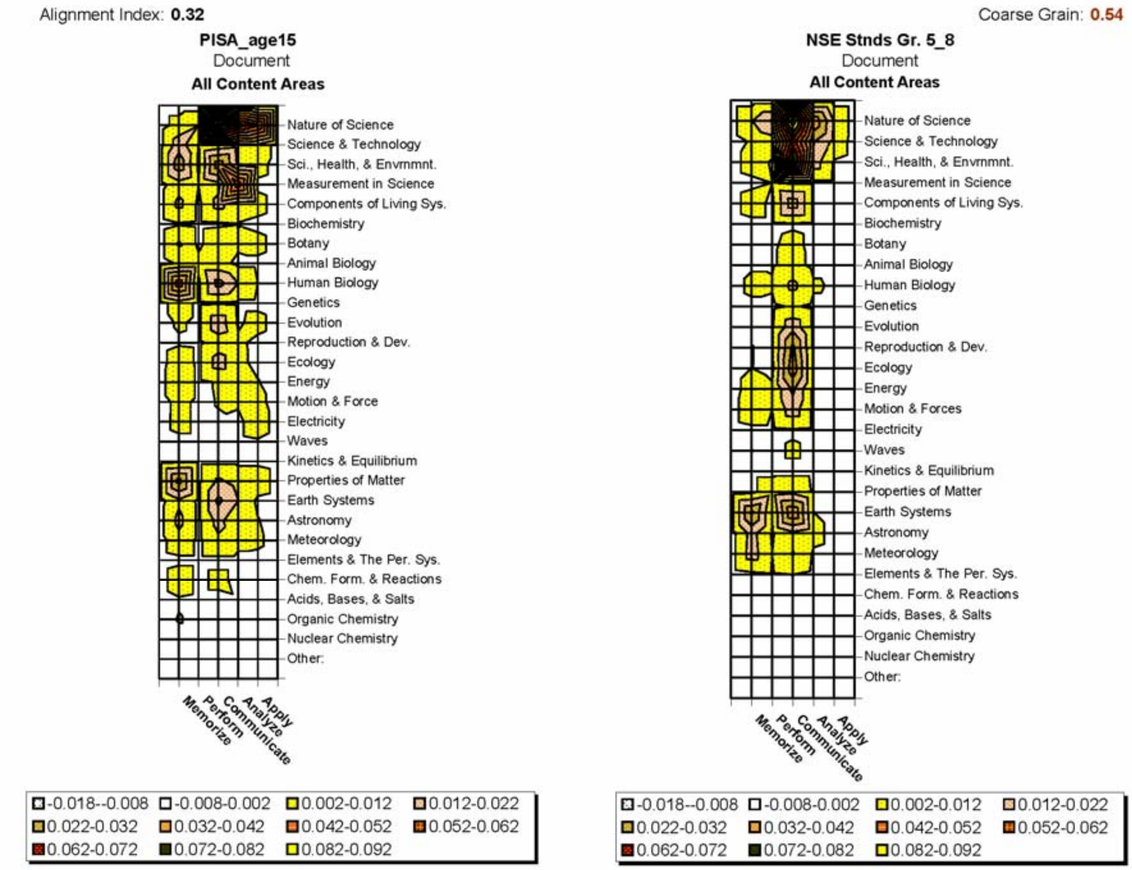
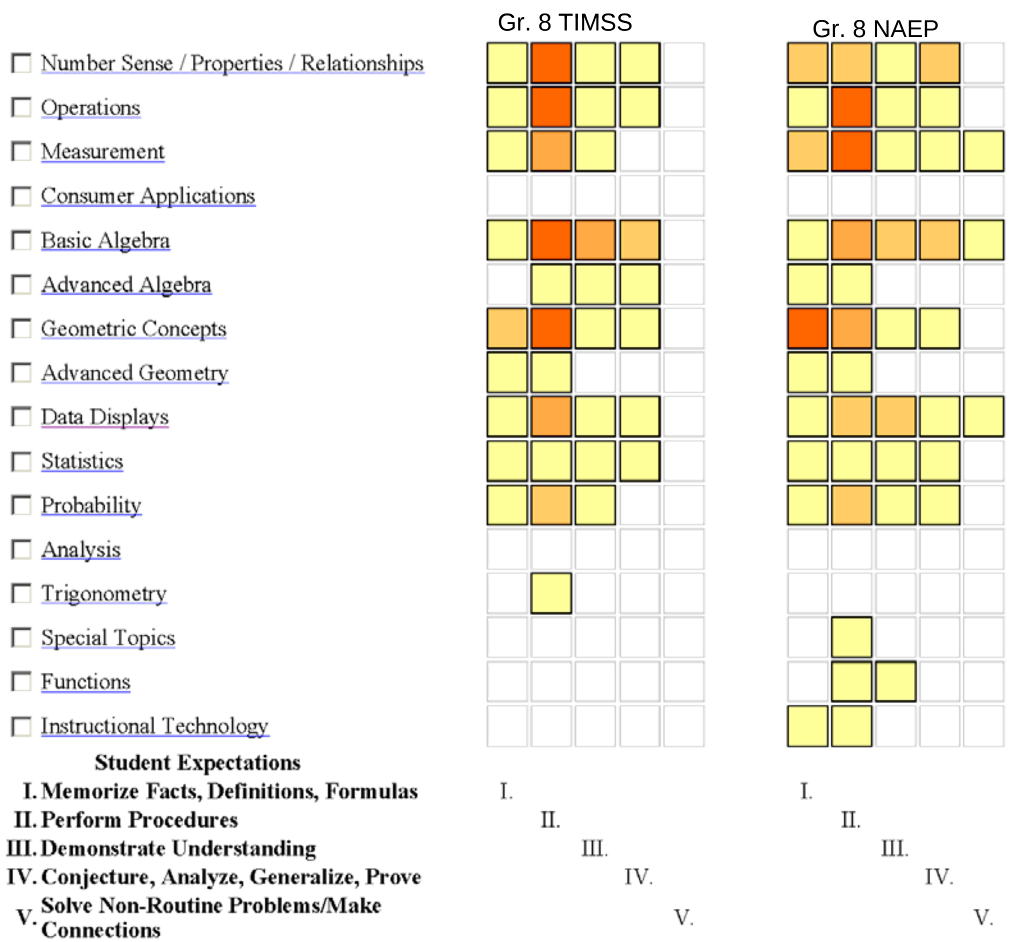


Figure 5 compares the PISA science item pool with the National Science Education Standards. One characteristic of the PISA assessment immediately noticeable is the amount of emphasis (i.e., proportion of assessment items) placed on topics related to the Nature of Science. This strong emphasis on Nature of Science helps to explain the strong alignment of the science PISA item pool (versus the TIMSS or NAEP item pools) to the National Science Education Standards. One noticeable, perhaps surprising, difference between the science PISA item pool and the National Science Education Standards is that PISA assesses recall knowledge on a broader range of science content than represented in the NSES, while also assessing more content at the apply/make connections level of performance expectation.

Coarse grain content descriptions of the Grade 8 TIMSS and NAEP item pools are presented in Figures 6 and 7 for comparison purposes. Figure 6 displays a tile chart format for mathematics, and Figure 7 uses a content map display. Note that the NAEP results reported in Figure 7 are based on released items from the 2000 and 2005 Grades 8 and 12 NAEP assessments. The entire NAEP science item pool is slated for content analysis later this year. Further, the NAEP released item results presented in Figure 7 are not available to the SEC Online system. Once the complete item pool has been analyzed, the NAEP science content results will be made available to the public through the SEC Online website.

Figure 6

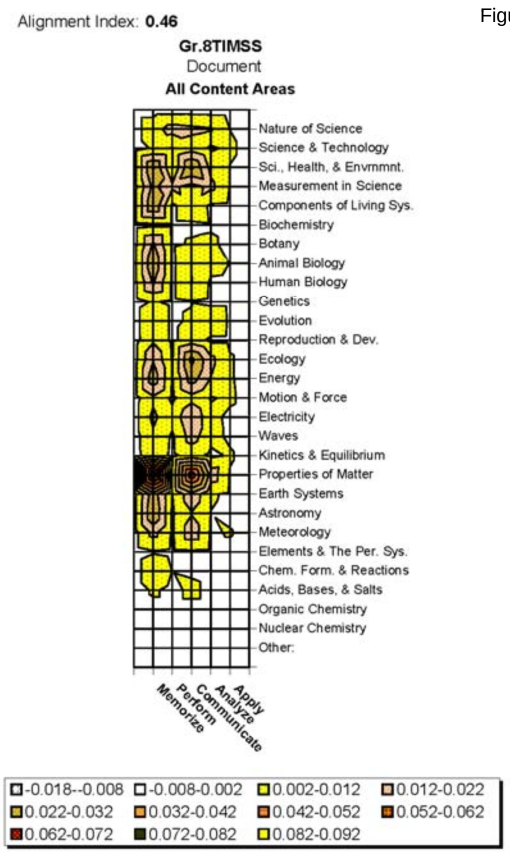
Mathematics Content: All Content Areas

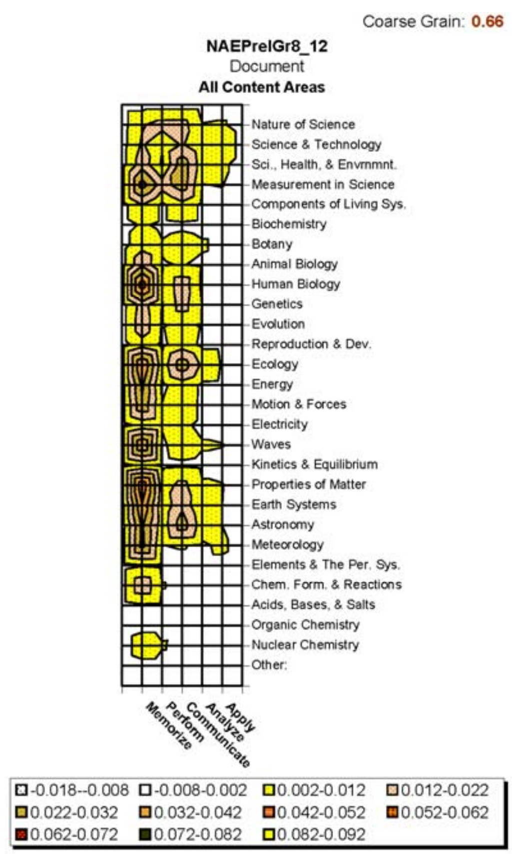


The content descriptions of Grade 8 TIMSS and Grade 8 NAEP item pools look quite similar, as can be seen in Figure 6. The most notable difference is that the Grade 8 NAEP item pool includes some items in the areas of Measurement, Basic Algebra, and Data Displays at the Solve Routine Problems/Make Connections performance level. By contrast, none of the items in the Grade 8 TIMSS item pool were designated by the content analysts as assessing science topics at this performance level.

In science as well, both Grade 8 NAEP and TIMSS items pools assess quite similar content. Two performance expectation categories are most represented in both items pools: Memorization and Communicate Understanding. However, all five categories of performance expectations are represented to some degree in both item pools.

Figure 7





SUMMARY

The SEC content analysis methodology provides a unique and detailed view of the academic content embedded in key target documents and item pools related to academic instruction. Results of content analyses serve both descriptive and analytic purposes. The content descriptions made available through tile charts and/or content maps provide a useful tool for teachers and other education stakeholders to review the relative amount and type of content embedded in key documents designed to either inform or assess instruction and learning.

These content descriptions also serve as inputs for alignment analyses that provide a highly quantitative and systematic means for comparing the similarity and differences between relevant instructional targets. By understanding the extent to which standards and assessments are related (aligned) to one another, educators and policymakers can make better informed decisions about the appropriateness of various assessments and other curriculum-related materials, for assessing and/or delivering standards-based content to all students.

Results of content analyses conducted on the PISA item pools for mathematics and science reveal an assessment that stands somewhat apart from the other high-profile assessments in mathematics and science, NAEP and TIMSS. Based on the review conducted here, the appropriateness of the content covered in the mathematics PISA item pool as a measure of U.S. mathematics proficiency remains a somewhat open question, largely because of the relatively low alignment to the NCTM standards. As a result, state education agencies are encouraged to compare the content assessed on the PISA with their own state standards to better determine the appropriateness of PISA assessments and results for their state. The SEC Online system provides a convenient resource for conducting such comparisons for those states that have had their standards analyzed previously with the SEC methodology.

In science, the PISA assessment appears to be quite well aligned to the National Science Education Standards, showing higher alignment to those standards than reported for either the Grade 8 TIMSS or NAEP. While state education agencies would be well served to similarly review the alignment of the PISA science item pools against state science standards (just as for math), the strong alignment between PISA and the NSES suggest a stronger case for keeping an eye on the PISA assessment results for science.

A key question for stakeholders and consumers of assessment results for both subject areas is the extent to which literacy, an important characteristic of the PISA assessment items, and the comprehension and writing skills thereby implied, represent a fair assessment of U.S. performance among 15-year-olds in science and mathematics.

REFERENCES

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).

Porter, A.C. & Smithson, J.L. (2001). Defining, developing, and using curriculum indicators. CPRE Research Report Series RR-048. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

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.