Athens Journal of Education - Volume 3, Issue 4 – Pages 313-330

What Do Mathematics Achievement Examinations Assess? A Critical Item Analysis

By Azita Manouchehri[*]
Xiangquan Yao[†]
Ali Fleming[‡]
Monelle Gomez[+]

In this work we examined the content of approximately 950 mathematics questions used in measuring achievement among 4th, 5th and 6th graders in the US in an attempt to determine how the current standardized examinations, which were developed to reflect new national curricular mandates, compared to the old measures used for the same purpose. Using Bloom's Taxonomy and TIMSS characterization of cognitive demands of tasks, we compared the quality of knowledge assessed by the items. Notable differences were found between the new items measuring skills and competencies compared to compatible items on the old tests. We noted that previous assessments primarily tested children's procedural and skill-based knowledge of mathematics, and often allowed children to arrive at a right answer in absence of understanding. In contrast, the new assessments require that children have a conceptual understanding of mathematics, attend to precise mathematical language, communicate their ideas, and utilize abstract reasoning skills.

Keywords: assessment, children's mathematical knowledge, mathematics, primary grades.

Introduction and Context

Lack of satisfaction with American students' mathematics achievement has a long history in mathematics education (Stanic & Kilpatrick, 2004). Calls for reform in how mathematics is taught and instruction organized in schools have been ongoing since the turn of the 20th century (Klein, 2000; Kilpatrick, 1992) however, published results of students' performance on international measures such as Program for International Students Assessment (PISA) and Trends in the International Mathematics and Science Study (TIMSS), on which American students were reportedly scored below the international average raised the urgency for improving teaching and learning of mathematics a national agenda. Indeed, poor performance of American children was for the

[*] Professor, The Ohio State University, USA.
[†] PhD Student, The Ohio State University, USA.
[‡] PhD Student, The Ohio State University, USA.
[+] PhD Student, The Ohio State University, USA.

https://doi.org/10.30958/aje.3-4-2 doi=10.30958/aje.3-4-2

most part associated with absence of a national curricula and lack of familiarity with the type of conceptual questions used on international measures (Scott, 2004; Provasnik, Lin, Darling, & Dodson, 2013). To address these two issues, in 2010, the United States introduced the Common Core State Standards for Mathematics (CCSSM), which outlines the content standards and mathematical dispositions to be achieved by children in schools across the country. While a national curriculum is not a new phenomenon for nations worldwide, the shift to a national curriculum is new in the United States and has brought about changes in the academic expectations and assessment practices of children. Prior to this initiative, each state created their own grade level content standards and administered assessments at each grade to reflect these standards. In other words, each state maintained different mathematical standards and assessment practices, which created scenarios in which children at the same grade level may have been learning and assessed on different content, and at different cognitive demands. This conflict posed problems for students competing for the same college admission positions, college scholarships, and positions in the job market. The introduction of the CCSSM national curriculum sought to address these issues by requiring children to think about and understand mathematics on a more conceptual level through the implementation of the common content standards.

With the publication of common standards, the movement towards developing new and common assessment tools followed. Children across the country would no longer take assessments specific to the state in which they lived; instead, children would take common grade-level specific national assessments created to reflect the mathematics content and problem solving practices outlined in the CCSSM. In this process, two assessment consortia were created (Partnership for College and Career Readiness and Smarter Balanced Assessment) to create national assessments that would reflect the new CCSSM standards. Each state could then choose the consortia that would provide the common assessments at each grade level across the state. Although the assessment of children's knowledge is just a snapshot of their growth, usually measured at the end of each school year, it is also known that what is tested is what is taught. In the transition from different sets of standards and assessments for each of the 50 states to a national set of standards and assessments at each grade level, changes will need to be made in classrooms across the country in order to see that these standards are enacted with fidelity.

While teachers will be held accountable for student results on these new assessments, teacher educators will need to design experiences for teachers to prepare them for helping children meet these expectations. To better understand the differences between the prior state assessments as compared to the new assessments, we evaluated approximately 800 items from prior state assessments in grades 4, 5, and 6, and 150 released items from the new national assessments in the same grades in an attempt to determine the type of knowledge these items elicited. The goal was to identify areas of emphasis, relative cognitive load of tasks and the type of teaching that they may demand.

Literature Review and Framework

The relationship between norm-based assessments, including student achievement examinations and instruction has long been discussed in the educational research (Madaus 1988; Au, 2007). There is understanding that the content of such measures often drive the quality and content of instruction that takes place in classrooms (Shepard & Dougherty, 1991; Clark, et.al., 2003; Polesel, Rice, & Dulfer, 2014). With stressed emphasis on teacher accountability based on learners' performance on tests, it is agreed that success in efforts to improve teaching and learning depends largely on aligning how and what knowledge is assessed (Madus, 1988).

Madaus (1988) offered a set of six principles that describe the consequences of measurement-driven instruction and show how they affect teacher and student behavior and the test itself. Three out of the six principles are discussion of the relation of teaching and testing. Those three principles are: i) if important decisions are presumed to be related to test results, then teachers will teach to the test; ii) in every setting where a high-stakes test operates, a tradition of past tests develops, which eventually de facto defines the curriculum; iii) teachers pay particular attention to the form and format of the questions on a high-stakes test and adjust their instruction accordingly (p.37-41).

To study the effects of standardized testing on instruction, Lorrie and Cutts (1991) surveyed 360 teachers from grade 3 to grade 6 to learn their test preparation/coaching practices and the effects of testing on instruction. The results from their survey indicated that teachers gave greater emphasis to basic skills instruction and teachers felt that content not tested suffered because of the focus on the standardized tests and testing further distorted teaching because of the extensive time given to test preparation.

Koretz, McCaffrey, and Hamilton (2001) identified seven types of teacher preparation to high-stakes tests, i.e. providing more instructional time, working harder to cover more material, working more effectively, reallocating classroom instructional time, aligning instruction with standards, coaching students to do better by focusing instruction on incidental aspects of the test, cheating (p.16). According to the author, the first three forms of teacher response have positive effects on student achievement; the next four forms have ambiguous effects on student achievement; and cheating is clearly negative.

Clarke and his colleagues (2003) surveyed teachers and education administrators in Kansas, Michigan and Massachusetts to understand their perceptions of state test on classroom practice and students. Results of the survey revealed that educators from the three states reported that preparing for the state test involved varying degrees of removing, emphasizing, and adding curriculum content and changes of teachers' instructional and assessment strategies. Perceived positive effects of these instructional changes included re-emphasis on writing, critical thinking skills, discussion, and explanation, while perceived negative effects included reduced instructional creativity, increased preparation for tests, a focus on breadth rather than depth of content coverage, and a curricular sequence and pace that were inappropriate for some students. In those states, only about 10% of interviewees felt that the state test did not affect instructional or assessment strategies.

Pedulla and colleagues (2003) distributed an 80-item survey to teachers across states to study their attitudes towards and opinions of state testing programs. They found a strong interaction between the level of stakes in the test and the degree to which it impacted curriculum and instruction. Across all types of testing programs, teachers reported increased time spent on subject areas that are tested and less time on areas not tested. Teachers also reported that testing has influenced the time spent using a variety of instructional methods such as whole-group instruction, individual-seat work, cooperative learning, and using problems similar to those on the test. Similarly, in a survey to study teachers' perspective on high-stakes test, Taylor and his colleagues (2003) found that teachers in Colorado reported added new content, such as probability and geometric topics, and emphasized more on problem solving and writing in mathematics classes, meanwhile decreased the amount of time spent on other curricular areas, such as science and social studies.

Polesel, Rice and Dulfer (2014) surveyed over 8000 Australian teachers to study their perspectives on their new national assessment program. Findings from the survey indicate that the testing regime is leading to a reduction in time spent on other curriculum areas and adjustment of pedagogical practice and curriculum content to mirror the tests.

Using the method of qualitative meta-synthesis, Au (2007) analyzed 49 studies on testing and curriculum and identified contradictory trends of the relationship between high-stakes testing and classroom practice. The major effect of high-stakes testing is that implemented curricular content is narrowed down to tested subjects, subject domain knowledge is fragmented into test-related pieces, and teacher-centered pedagogies become prevalent in classroom. However, Au also found that certain types of high-stakes tests have led to curricular content extension, the integration of knowledge, and more student-centered instructional strategies. The contradictory trends suggest that the structure of a test mediates the relation of standardized test and classroom practice. This is evident both locally and internationally (Scott, 2004).

Calls for reforming school mathematics remain futile if teachers remain unclear about the quality of knowledge they are expected to nurture. Moreover, enhancement of performance on the part of the learners remains an illusive goal if how knowledge is assessed is not reflective of what knowledge is to be taught. The desire to address these goals motivated the research reported here.

Methodology

The goal of our study was to first determine and then to compare the content of new (reform based) and old standardized achievement examinations used in the US. Our research followed an emergent model associated with textual analysis (Strauss & Corbin 1990) in which the purpose of analysis aimed to capture patterns and regularities in texts (Silverman 2005, 2006).

For comparison, we selected released standardized achievement tests designed for grades 4, 5, and 6 from 6 different states (California, Texas, Minnesota, Massachusetts, New York, Ohio). In order to establish whether the state tests were consistent in what and how they assessed children's mathematical knowledge, two consecutive tests at each grade level were investigated from Ohio and Massachusetts, while one test at each grade level was studied for the remaining four states. For the purpose of our analysis, we considered the three largest states in the country, indicating the largest number of students that were tested (New York, California, Texas), and states with the highest student performance based on national and international rankings (Ohio, Minnesota, Massachusetts). We also acknowledge that the released national assessment items were not full exams at each grade level, but instead were released practice items.

All of the state items and the items from the new national assessments at grades 4, 5, and 6 were first cataloged by content strand. We then rated each of the items according to the type of thinking they required of students. In doing so, we utilized two different rating frameworks: Bloom's taxonomy and TIMSS Cognitive Domains. The decision to choose two complementary frames was to increase the precision of our analysis of the tasks and the cognitive skills they demanded. Since one of the major claims regarding the new assessment items is their compatibility of test items with those used on the international achievement examinations the use of TIMSS scoring rubric became most useful.

Rating Criteria

Each item was first classified according to one of five content strands: number sense/operations, measurement, geometry, algebra/pattern, and data analysis/ probability. Released content blueprints for each assessment studied were used to classify each item, in addition to the judgements of the four researchers.

Each item was then rated using Bloom's Taxonomy (Clark, 2013) and the Mathematics Cognitive Domains proposed by TIMSS (Grønmo, Lindquist, Arora, & Mullis, 2013). Bloom's taxonomy was chosen because of its widespread use in educational settings. The TIMSS framework was utilized as a way of framing the results using an international rating index so to allow for future and additional comparison of items. A brief description of each framework is offered below.

Bloom's Taxonomy

Bloom's taxonomy proposes six types of intellectual domains associated with knowing of any concept. These domains include Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. The *Knowledge* domain (B1) consists of recall of data or information (Clark, 2013), ranging from specific facts to complete theories. *Comprehension* (B2) involves understanding the meaning of concepts and the ability to translate material from one form to another (words to numbers, graphs to symbols). The *Application* level (B3) encompasses the use of concepts in a new and unfamiliar context relying on a higher level of understanding compared to Comprehension. *Analysis* (B4) assumes the individual is capable of identifying facts and inferences drawing from structural knowledge. *Synthesis* (B5) demands creating new meaning, or building a structure or pattern. At this level the individual can put together a set of abstract relationships to classify objects (Clark, 2013). *Evaluation* (B6) involves the ability to judge the value of statements, examine accuracy of conjectures, and form propositions supported by evidence.

TIMSS Cognitive Domains

The TIMSS Cognitive Domains consist of three broad levels including: Knowing, Applying, and Reasoning (Grønmo, Lindquist, Arora, & Mullis, 2013). The *Knowing* domain consists of knowledge of facts, algorithms, and theorems. In this domain the individual is capable of performing tasks in contexts that are familiar, regardless of the level of difficulty of the task itself. *Applying* involves problem solving and implementing strategies that require extension of known facts and algorithms. The *Reasoning* domain includes the ability to form conclusions, connect multiple representations of concepts, formulate generalizations, and identify structural connections.

Four independent researchers evaluated each of the items for content strand and both cognitive demand frameworks, and their individual rankings were then compared. In places where the raters disagreed on the ranking of a problem, the item was extensively discussed until consensus was reached. When failure to reach consensus on the cognitive demand ranking of an item occurred, the task was granted the highest ranking awarded by the team members. Twenty items (approximately 2% of the total items considered) fell into this category. The items were equally distributed among all tests and did not alter the results in the final analysis.

Results

Table 1 summarizes the percentage of items in each of the content strands included in the released state assessments and the new national practice assessments in grades 4, 5, and 6. It is important to note here that the 150 items from the new national assessment were not from a complete exam, but rather released practice items at each grade level. The national items also did not include performance-based assessment items, which will part of the common assessments when administered in classrooms. Therefore, the distribution of items according to content strands is not complete and may not be an accurate representation of what is tested.

Significant differences in content strand were noted between prior state assessments and the new national assessments at grades 4 and 5, while grade 6 exams were compatible. While the majority of the items in all the tests and at all grade levels pertained to the Number Sense/Operations strand, this percentage decreased with each grade level, as the percentage of Algebra/ Pattern items generally increased.

Table 1. Frequency and Percentage of State and National Items by Grade and Content Strand

		Number of items	**Number Sense/ Operations**	**Measurement**	**Geometry**	**Algebra/ Pattern**	**Data Analysis/ Probability**
Grade 4	*State Avg.*	48	23 (48.4%)	7 (14.2%)	6 (12.8%)	7 (15.9%)	5 (9.7%)
	National	44	35 (79.5%)	6 (13.6%)	3 (6.8%)	0	0
Grade 5	*State Avg.*	46	19 (40.1%)	10 (21.3%)	6 (12.6%)	8 (17.3%)	4 (8.7%)
	National	44	28 (63.6%)	9 (20.5%)	6 (13.6%)	1 (2.3%)	0
Grade 6	*State Avg.*	45	20 (43.8%)	7 (16.2%)	3 (6.6%)	8 (18.0%)	7 (15.4%)
	National	42	20 (47.6%)	6 (14.3%)	4 (9.5%)	7 (16.7%)	5 (11.9%)

In grade 4, the major differences between the prior state exams and the new national items concerned the average percentage of items that measured knowledge of Geometry, Data Analysis/Probability, and Algebra/Pattern. Indeed, there were no released items pertaining to Data Analysis/Probability and Algebra/Pattern on the national assessment for grade 4. The national assessment instead placed greater emphasis on Number Sense/Operations, with 79.5% of the items measuring children’s knowledge in that domain. This is a significant increase compared to the state average of 47.3% of the items in this domain. At the fourth grade, the distribution of items that focused on measurement was equivalent in the prior state assessments and the new national assessment.

In grade 5, the largest discrepancy concerned the percentage of the items that measured the Algebra/Pattern domain. Approximately 64% of the items on the new national assessment measured Number Sense/Operations and included only one item that concerned Algebra/Pattern. In contrast, the state average percentage of Algebra/Pattern items reached 17.4%. The new national practice assessment items did not include items that measured stochastic reasoning in fifth grade.

Overall, as the grade level increased, the emphasis on algebraic reasoning also increased, with a large percentage of items addressing the Number

Sense/Operations strand at each grade level. The largest differences were visible at the fifth grade where Geometry knowledge was most prominently emphasized and demanded the most sophisticated type of reasoning. This latter issue will be discussed later in this paper.

What Cognitive Demand Do The New National Items Require? A Quantitative Comparison

Table 2 summarizes the percentage of items according to the type of knowledge and level of understanding these items tend to elicit from children, as well as the overall mean ranking of released items on Bloom's and TIMSS levels. At each grade level, the state assessments are compared with those of the new national assessments. We calculated Pearson's χ^2, with the null hypothesis assuming that the state items had the same Bloom's level and TIMSS level distribution as the released new national assessment items. Each χ^2 statistic was compared to a chi-squared random variable using a significance level of α=0.05. χ^2 tests indicate that there is no statistically significant difference between the prior state assessments and the new national assessments on TIMSS level and Bloom's level. Note, though, the larger percentage of national items that are ranked at Bloom's levels 3 and 4 as compared to the state averages, particularly at grade 5.

Table 2. Percentage of State and National Items by TIMSS and Bloom's Level

		Number of items	**TIMSS level**			**Bloom's level**				
			1	*2*	*3*	*1*	*2*	*3*	*4*	*5*
Grade 4	*State*	289	52.9	39.4	7.6	32.5	48.8	14.2	4.5	0.0
	National	44	54.5	40.9	4.5	36.4	40.9	20.5	2.3	0.0
		χ^2=0.537 P=0.764				χ^2=2.706 P=0.439				
Grade 5	*State*	277	49.5	40.8	9.7	35.4	46.2	12.3	5.8	0.4
	National	44	47.7	38.6	13.6	34.1	31.8	22.7	11.4	0.0
		χ^2=0.625 P=0.731				χ^2=6.335 P=0.096				
Grade 6	*State*	272	41.9	45.6	12.5	26.8	40.8	24.6	7.4	0.4
	National	42	35.7	52.4	11.9	23.8	40.5	23.8	11.9	0.0
		χ^2=0.712 P=0.701				χ^2=1.087 P=0.780				

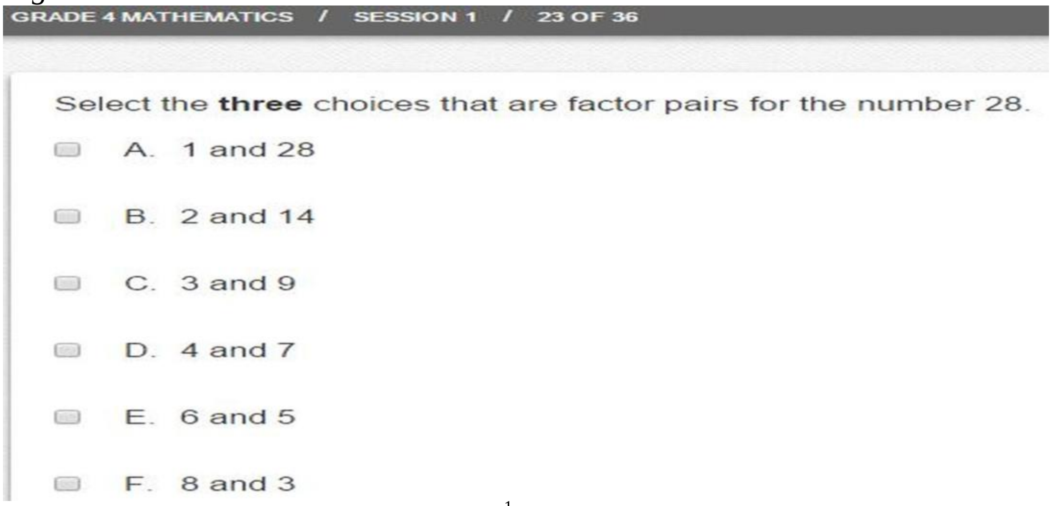
Discussion

Although our quantitative analysis showed that the new assessment items and those used on prior state achievement exams were compatible according to Bloom's and TIMSS cognitive demand levels, a close inspection of the content of the new national items revealed significant differences between them and those of the prior state examinations. Aside from demanding greater organizational skills, items on the new tests require greater attention to and proficiency with mathematical language, decision-making ability, abstract reasoning skills, and modeling with mathematics. In the following section we will offer illustrative examples to highlight each of these differences. We chose to include items from the fourth and fifth grade tests only to emphasize the expectations the items place on children's mathematical thinking skills, some of which have traditionally been assumed to occur in higher grade levels.

Multiple Correct Responses vs. One Single Answer

While a majority of the prior state items required selecting one answer from a set of options, the new national items often asked students to select multiple correct answers from a list of choices given. This type of questioning requires that children be sensitive to various properties of the same concept or different forms the concept may take. A fourth grade question is presented in Figure 1. Notice that even though the items measures knowledge of facts, it demands a more complete response from children. Similar problems on prior state exams asked students to identify only one factor or one pair of factors of a given number, frequently using numbers with only two prime factors. Similarly, a fifth grade question from the new national assessments asked students to choose multiple correct statements from a list of 5 options when describing the coordinate system.

Figure 1. National Grade 4 Item #23



GRADE 4 MATHEMATICS / SESSION 1 / 23 OF 36

Select the **three** choices that are factor pairs for the number 28.

□ A. 1 and 28

□ B. 2 and 14

□ C. 3 and 9

□ D. 4 and 7

□ E. 6 and 5

□ F. 8 and 3

Source: Retrieved from http://bit.ly/1UwswHf [1]

Perform the Operation vs. Select the Right Answer

The new national test items, at all grade levels, even when they measured knowledge of facts and procedures, for the most part, asked students to perform operations and then provide an answer. In the state exams, since a majority of the items were multiple choice items, children could select the right answer without knowing the concept because they could use a guess-and-check

[1] To access items online the user needs to sign in to the site.

strategy to mark the correct answer. Prior state test items frequently allowed students to acquire a correct answer without an understanding of concepts since multiple choice items provided students with visual clues for how an acceptable answer may look while eliminating wrong answers resulting from minor computation errors. The new national practice items on the other hand, required students to demonstrate proficiency in carrying out the procedures.

Table 5 summarizes the percentage of multiple choice and open-ended responses among the pool of items considered in our analysis. Although this distribution is skewed since some of the state practice tests did not include samples of their open response items, it is evident that the new national assessment items tend to measure performance and understanding more heavily than state tests.

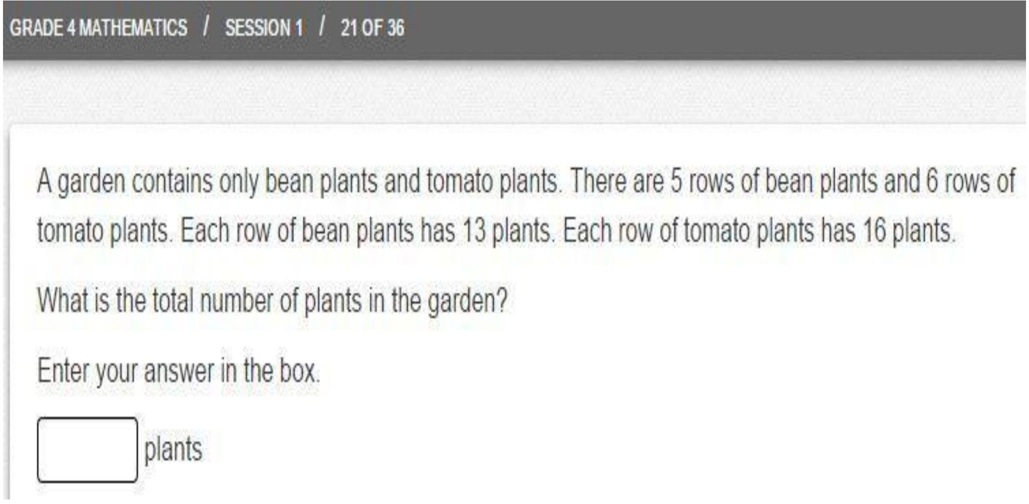
Table 3. Number and Percentage of the Types of State and National Assessment Items

State	**Grade 4**		**Grade 5**		**Grade 6**	
	Multiple Choice	*Open Response*	*Multiple Choice*	*Open Response*	*Multiple Choice*	*Open Response*
State Item	247 (85.5%)	42 (14.5%)	240 (86.6%)	37 (13.4%)	234 (86.0%)	38 (24.0%)
National Items	6 (13.6%)	38 (86.4%)	8 (18.2%)	36 (81.8%)	7 (16.7%)	35 (83.3%)

Multiple Step Computations vs. Single Step Computations

Although all exams contained knowledge of facts and a large quantity of application tasks, the new national exam items tended to present problems whose solutions required that the students go through organizing data first and then conduct multiple steps to obtain a result. The state achievement tests included, for the most part, problems that demanded a single step to reach an answer. Figure 2 displays an example of a typical application problem that the national assessments included. Notice that in this example, the children are expected to keep track of values, multiply different pairs of numbers, and then combine the results. Compatible items on the state exams that measured the same mathematical skill included only two values (i.e. A garden contains 5 rows of tomato plants, each row containing 7 plants. What is the total number of plants in the garden?).

Figure 2. National Grade 4 Item #21



GRADE 4 MATHEMATICS / SESSION 1 / 21 OF 36

A garden contains only bean plants and tomato plants. There are 5 rows of bean plants and 6 rows of tomato plants. Each row of bean plants has 13 plants. Each row of tomato plants has 16 plants.

What is the total number of plants in the garden?

Enter your answer in the box.

□ plants

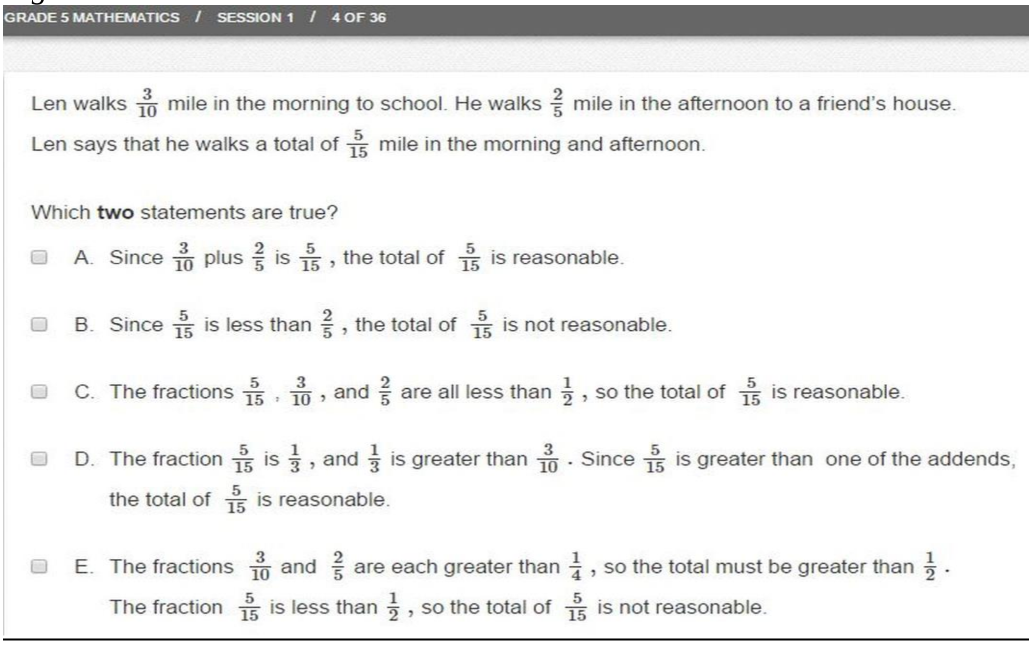
Source: Retrieved from bit.ly/1SchdAN.

Also of note was that even tasks that tested the same content area and secured the same Bloom's and TIMSS ratings still required different levels of cognitive demand and mathematical understanding. For example, one grade 5 item from the new national assessments asks children to find the answer to $\frac{3}{4}+\frac{4}{5}-\frac{7}{10}$ (retrieved from http://bit.ly/1MiUgNa). State items measuring the same content (fraction operations) required only one operation and had fractions with either common denominators or denominators that were relatively prime. For example, a grade 5 state test item asks students to find the equivalent expression to $\frac{2}{5}+\frac{1}{4}$ (retrieved from http://bit.ly/22rCpr8). Similarly, another grade 5 item from the state assessments places fraction operations within a context, but common denominators are given (Randa ate 3/8 of a pizza, and Marvin ate 1/8 of the same pizza. What fraction of the pizza did Randa and Marvin eat?) (retrieved from http://bit.ly/1UeOxe4). Additionally, the state items are both multiple choice, while the national item is an open response item.

Precise Use of Language

The new national assessment practice items contained more questions that demanded precise use of mathematical language than state items. Indeed, the demand for understanding and using mathematical language was central to successful completion of nearly 50% of all national assessment items. These items tended to be more sensitive to precision in describing numbers depending on contexts used. For instance, questions that asked students to locate points on a number line referenced exclusively fractions as rational numbers. Figure 3 offers an example of a problem typical of the quality and quantity of reading involved in some of the national items, along with the heavy emphasis on understanding of and sensitivity to mathematical language. Note that the item measures a number of skills in concert including ordering of fractions, addition of unlike fractions, reasoning based on magnitude of fractions, and determining reasonableness of answers. Compatible items which measured knowledge of fractions on state tests merely asked students to either estimate, locate, or order fractions on a number line.

Figure 3. National Grade 5 Item #4



GRADE 5 MATHEMATICS / SESSION 1 / 4 OF 36

Len walks $\frac{3}{10}$ mile in the morning to school. He walks $\frac{2}{5}$ mile in the afternoon to a friend's house.

Len says that he walks a total of $\frac{5}{15}$ mile in the morning and afternoon.

Which **two** statements are true?

□ A. Since $\frac{3}{10}$ plus $\frac{2}{5}$ is $\frac{5}{15}$, the total of $\frac{5}{15}$ is reasonable.

□ B. Since $\frac{5}{15}$ is less than $\frac{2}{5}$, the total of $\frac{5}{15}$ is not reasonable.

□ C. The fractions $\frac{5}{15}$, $\frac{3}{10}$, and $\frac{2}{5}$ are all less than $\frac{1}{2}$, so the total of $\frac{5}{15}$ is reasonable.

□ D. The fraction $\frac{5}{15}$ is $\frac{1}{3}$, and $\frac{1}{3}$ is greater than $\frac{3}{10}$. Since $\frac{5}{15}$ is greater than one of the addends, the total of $\frac{5}{15}$ is reasonable.

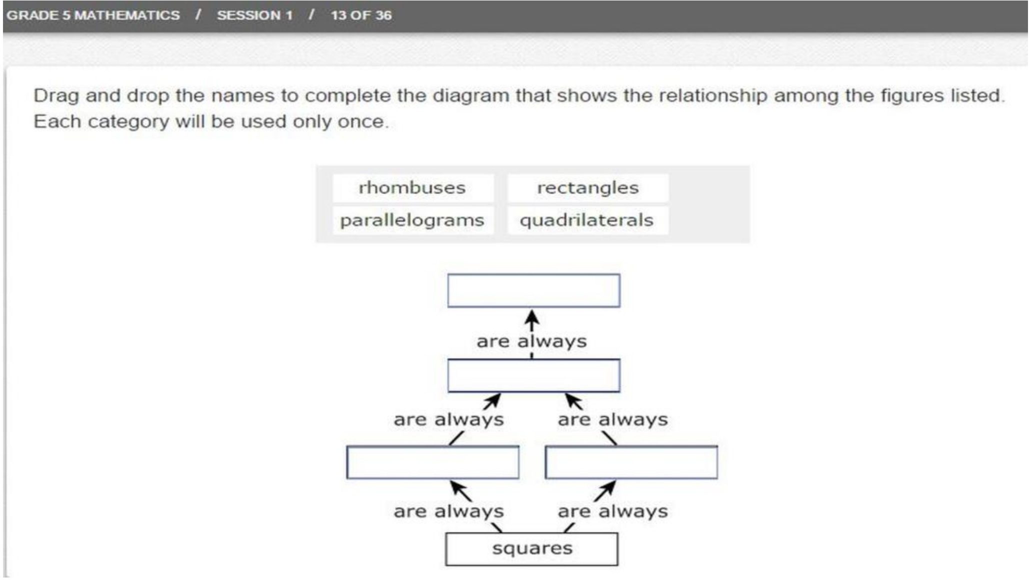
□ E. The fractions $\frac{3}{10}$ and $\frac{2}{5}$ are each greater than $\frac{1}{4}$, so the total must be greater than $\frac{1}{2}$. The fraction $\frac{5}{15}$ is less than $\frac{1}{2}$, so the total of $\frac{5}{15}$ is not reasonable.

Source: Retrieved from http://bit.ly/1Rbzaix.

Structural Knowledge vs. Defining Properties

Common among all reasoning items on the new national released items, regardless of the grade level or content strand, appeared to be the demand placed on learners to consider and formulate generalizations by taking into account class relationship. This phenomenon was almost profoundly visible in the context of geometry items and most prominent on the fifth grade practice test. The example in Figure 4 is typical of this group of tasks. Notice that although all state tests included geometry tasks that asked students to identify one or two properties of quadrilaterals, the expectation of knowledge was primarily at Level 1 (Knowledge) of Bloom's Taxonomy. In the example in Figure 4, the children need to know not only the properties of objects as discrete entities but also the hierarchical relationship among them. The problem elicits the kind of thinking rarely emphasized in school curriculum. (Note: Although the direction of arrows and/or the quadrilateral options the problem provides may have been the misfortunate result of a typographical error, the type of thinking the problem demands remains not only mathematically valuate, but also novel in school curriculum).

Figure 4. National Grade 5 Item #13



GRADE 5 MATHEMATICS / SESSION 1 / 13 OF 36

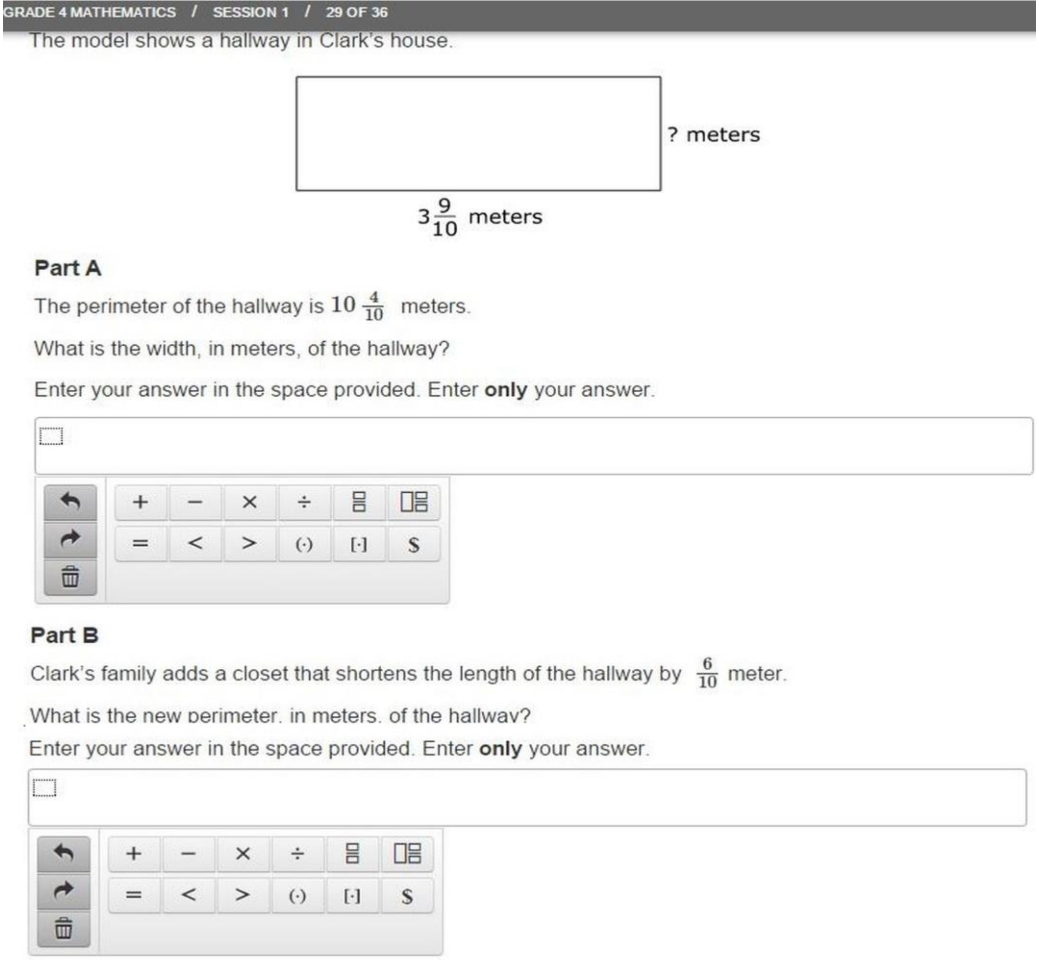
Drag and drop the names to complete the diagram that shows the relationship among the figures listed. Each category will be used only once.

Source: Retrieved from http://bit.ly/1WBGC8b.

Abstract Reasoning vs. Extracting Clues from Visual Imageries

Although in the previous section we offered a description of the type of tasks that demanded abstract and generalized thinking, a closely linked major difference between the new national items at all grade levels and those on previous state exams included the level of abstraction and formal knowledge expected of children when solving problems. Although the state items often included images, representations, pictures, and models from which the children could extract information to solve problems, the new national items either did not provide such visual clues or when they did, the images could not be used directly to find answers. On state items, most typically, when children were asked to find the area, perimeter, or volume of geometric shapes, those shapes were presented on a grid where children could count to obtain answers. Additionally, whole numbers were often used rather than fractions. Such was not the case for similar items on the new national assessments. Notice that in the example showing in Figure 5, children need to alter or manipulate the accompanied image to answer the question. Notice also that the visual media presented to children does not include features that could be used to solve the problem without understanding the concept of perimeter. Indeed, knowing the formula alone would not be sufficient to solve the tasks.

Figure 5. National Grade 4 Item #29



GRADE 4 MATHEMATICS / SESSION 1 / 29 OF 36

The model shows a hallway in Clark's house.

Part A

The perimeter of the hallway is $10\frac{4}{10}$ meters.

What is the width, in meters, of the hallway?

Enter your answer in the space provided. Enter **only** your answer.

Part B

Clark's family adds a closet that shortens the length of the hallway by $\frac{6}{10}$ meter.

What is the new perimeter, in meters, of the hallway?

Enter your answer in the space provided. Enter **only** your answer.

Source: Retrieved from bit.ly/22rtSrM.

We do stress that the analysis offered in this paper was based solely on the available practice items from one of the two assessment consortia creating the new national assessments, and on items from released state sample tests. The released tests did not include all open response items or performance-based tasks that are used. Any inferences about the state tests or national assessments must be cautionary due to the fact that neither full examinations nor an accurate distribution of items according to content strands was made available. The primary value of analysis is to help identify new types of knowledge children would need to master on the new national assessments. It is often said that in educational settings what is tested is what is taught. If that is indeed the case, and if the released national practice items are a representative of what knowledge is to be valued, then there is some indication that teachers must help children develop skills consistent with mathematical practices so to successfully meet the demands of the new assessments. As discussed in previous sections, while no significant differences existed among the achievement tests used by different states and the new national released achievement items according to the cognitive load they demanded of children, the degree of decision-making and abstract reasoning expected of children was remarkably higher than the standardized state achievement exams. Certainly both in quality and in quantity, the questions that measured abstract reasoning, structural knowledge, and tending to precision ranked higher on the new national items as compared to the prior state items.

As we mentioned earlier, questions addressing some of the content strands were not included among the items designed for certain grade levels on the new national assessment items, or they appeared less frequently. For instance, Geometry items were limited to only 4 in the fourth grade. However, in reviewing fifth grade Geometry items it is clear that meeting the mathematical expectations of knowledge of Geometry at that grade may not be met without adequate preparation in the previous years. The same applies to the Data Analysis/Probability items appearing in Grade 6.

A large number of the new national assessment items require students to do a substantial degree of abstracting. They also demand that students be flexible problem solvers. The skills needed for student success on these items require the skills to be built over time and through the use of activities and lessons that focus on the development of mathematical thinking among children. Attempts at educating mathematics teachers must capitalize on the demands of these tasks and engage them in explicit discussions focused on understanding the mathematical connections that need to be unveiled in instruction.

Final Comments

Early in this report we highlighted the importance of content analysis of tests that claim to measure mathematical thinking capacity of school children as a means to define venues for improvement of teaching. Although our work concerned on nationally used achievement tests issues pertaining to what and how knowledge of mathematics is elicited is of international concern. Over a decade ago Scott (2004) compared the design, features, framework, and items of Trends in International Mathematics and Science Study (TIMSS) Assessment Items, Program for International Students Assessment (PISA), and the National Assessment of Educational Progress (NAEP). They conclude that all the three assessments provide different lenses to view and better understand student performance. He argued for the need to extend such line of inquiry in order to develop a better understanding of not only the students' performance but also results associated with differences in achievement. In that same spirit, Grønmo and Olsen (2006) examined the content covered in the Trends in the International Mathematics and Science Study (TIMSS) (Grade 8 level) and the Program for International Students Assessment (PISA) and found the framework of the two assessments are based on two different perspectives of mathematics (*applied mathematics* vs. *pure mathematics*). According the authors, PISA emphasizes more the contexts and phenomena where mathematics competencies can be used in real world, while TIMSS gives much more attention the structure and formal aspects of mathematics. Findings of our research highlight a critical issue pertaining to the body of work that concerns item analysis of various examinations designed for measuring school learners' mathematical achievement either locally or globally. We found classification of the cognitive loads of the items based on the two analytical tools we used to be problematic. From the point of view of mathematical rigor, the levels in both models were inclusive of multiple stages of thinking. This clearly interrupted their ability to distinguish items according to the quality of knowledge they measured. Indeed, in the absence of our careful qualitative analysis of items it would be sensible to conclude that the new and old items were compatible, a false conclusion. This suggests that a mixed method of analysis, taking into account disciplinary knowledge might be essential if such work is to provide concrete guide for how teaching might be adjusted to accommodate for greater achievement based on tests.

References

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258-267.

Clark, D. (2013). *Bloom's taxonomy of learning domains*. Retrieved from bit.ly/1VtAE rD.

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from Interviews with Educators in Low-, Medium-, and High-Stakes States*. Boston: Lynch School of Education, National Board of Educational Testing and Public Policy.

Klein D. (2000). Math Problems: Why the U.S. Department of Education's recommended math programs don't add up. *American School Board Journal*, *187*(4), 52-57.

Grønmo, L. S., Lindquist, M., Arora, A., & Mullis, I. V. S. (2013). TIMSS 2015 mathematics framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS* 2015 Assessment Frameworks (Chapter 1). Retrieved from bit.ly/1TTJ8c5.

Grønmo, L. S., & Olsen, R. V. (2006). TIMSS versus PISA: The case of pure and applied mathematics. In *2nd IEA International Research Conference*.

Kilpatrick, J. (1992). A history of research in mathematics education. In D. A. Grouws, (ed.), *Handbook of Research on Mathematics Teaching and Learning*. New York: Macmillan.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). T*oward a framework for validating gains under high-stakes conditions*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29-46.

Massachusetts Department of Elementary and Secondary Education (2012). *Massachusetts Comprehensive Assessment System*. Retrieved from bit.ly/22r Cpr8.

Partnership for Assessment of Readiness for College and Careers (2014a). *Mathematics Practice Tests*. Retrieved from http://bit.ly/1dPG77N

Partnership for Assessment of Readiness for College and Careers (2014b). *PARCC field test*. Retrieved from http://bit.ly/1WBIKg7.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston: Lynch School of Education, National Board on Educational Testing and Public Policy.

Polesel, J., Rice, S., & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: a teacher perspective from Australia. *Journal of Education Policy*, *29*(5), 640-657.

Provasnik, S., Lin C., Darling, D., & Dodson, J. (2013). A Comparison of the 2011 Trends in International Mathematics and Science Study (TIMSS) Assessment Items and the 2011 National Assessment of Educational Progress (NAEP) Frameworks. *National Center for Education Statistics*.

Scott, E. (2004). Comparing NAEP, TIMSS and PISA in Mathematics and Science. *National Center for Education Statistics*.

Shepard, L. A., & Dougherty, K. C. (1991). *Effects of High-Stakes Testing on Instruction*. Paper presented at the annual meeting of the American Educational Research Association. Chicago.

Silverman, D. (2005). *Doing Qualitative Research: A Practical Handbook.* London: Sage.

Silverman, D. (2006). *Interpreting qualitative data: Methods for analyzing talk, text and interaction*. London: Sage.

Stanic, G. M., & Kilpatrick, J. (2004). *Mathematics curriculum reform in the United States: a historical perspective*. *Educação Matemática Pesquisa, 6*(2), 11-27.

Strauss, A. L., & Corbin, J. (1990). *Basics of Qualitative Research: Techniques and procedures for developing grounded theory*. Newbury Park: Sage.

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2003). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost* (CSE Technical Report 588). Los Angeles: CRESST.

U. S. Census Bureau (2014). *The 2012 statistical abstract*. Retrieved from http://1.usa.gov/1LAmHWL.

University of the State of New York - New York State Education Department (2010). *Regents exams*. Retrieved from http://bit.ly/1U0tlrY.