

Chapter 2

The Impact of Reading Ability on TIMSS Mathematics and Science Achievement at the Fourth Grade: An Analysis by Item Reading Demands

Ina V.S. Mullis, Michael O. Martin, and Pierre Foy
Boston College

Introduction

In the past several decades, schools have seen increasing integration across subject areas in teaching and learning, including greater emphasis on reading within subject areas. Today mathematics and science curricula around the world, as well as standards for proficiency in these subjects, commonly include reading and communication skills, and the *TIMSS 2011 Assessment Frameworks* (Mullis, Martin, Ruddock, O’Sullivan, & Preuschoff, 2009) reflect this situation. At the fourth grade in mathematics, for example, topics in the TIMSS framework content domains of number, measurement, and data display

specify that students should be able to solve routine and non-routine problems set in everyday contexts. Understanding the description of the everyday situations for these types of problems necessarily involves reading. Also, the data display area is based on “reading and interpreting” tables, pictographs, bar graphs, and pie charts as well as creating such data displays. Similarly, the science framework requires students to comprehend descriptions of experiments and investigations as well as to read and interpret a variety of models and diagrams of science systems and phenomena.

Given that reading has been incorporated into the mathematics and science assessment frameworks, the TIMSS 2011 fourth grade achievement items encompass a range of reading comprehension demands. Of course, the reading demands vary across items, from quite minimal, as in items requiring students to complete a calculation, to somewhat more substantial, as in items requiring students to understand a description of a science experiment or phenomenon and then apply their knowledge or explain their reasoning. Additionally, it should be emphasized that all TIMSS items undergo extensive review for clarity, straightforward vocabulary, and syntax, and that any extraneous or irrelevant information is removed. Developing clearly written items that are equally accessible to all students is fundamental to all item development, but especially pertinent to TIMSS which needs to be translated into 30 or so languages. Multiple reviews by representatives of the participating countries ensure that the TIMSS item development approach prioritizes clarity and brevity, such that none of the TIMSS fourth grade items involve reading of any length or complexity (although some do at the eighth grade).

Still, fourth grade students are likely to be at a disadvantage in learning mathematics and science as well as demonstrating high performance on the TIMSS assessment if they lack reading skills. The availability of PIRLS data on reading achievement at the fourth grade provides an ideal opportunity to investigate the relationship between reading ability and the reading demands of the TIMSS mathematics and science assessment items. This study capitalized on the unique availability of PIRLS and TIMSS achievement scores for the same fourth grade students across reading, mathematics, and science for 34 countries and three benchmarking participants to examine the following overarching question:

How does reading ability impact TIMSS mathematics and science achievement at the fourth grade?

Overview of Study

The basic approach used in the analysis presented in this chapter was to examine, for each participating country and benchmarking entity, the relationship between fourth grade students' reading ability as measured by PIRLS and their performance on TIMSS items with increasing levels of reading demands.

The hypotheses were as follows:

1. Students with high reading ability would not be impacted by the level of reading demand in the items. That is, the best readers would score similarly on TIMSS items regardless of the degree of reading demands.
2. Students with lower reading ability would perform relatively better on items with less reading. That is, poorer readers would score better on the items with the lowest reading demands than on the items with the highest reading demands.

The study was conducted separately for mathematics and for science, with the initial steps involving sorting the 175 TIMSS fourth grade mathematics items and 168 science items according to degree of reading demands. To maintain robustness of measurement while at the same time preserving differentiation, each set of fourth grade TIMSS items (mathematics and science) was separated into three relatively equal sized categories from relatively low to relatively high reading demands (low, medium, and high). Students' mathematics and science achievement was examined for each of the groups of mathematics and science items categorized as having low, medium, and high reading demands, for students at three different levels of reading ability on PIRLS 2011 (upper, middle, and lower terciles). The hypotheses were supported in general, but more so in mathematics than in science, and more so in some countries than in others. The results differed across countries, and sometimes between mathematics and science within countries.

Categorizing the TIMSS Fourth Grade Mathematics and Science Items According to Reading Demands

It was fundamental to this study to be able rank the TIMSS fourth grade mathematics and science items by level of reading demand in a manner that would be reliable and appropriate for further analysis. To have enough categories to discriminate between items, but not too many categories such that distinguishing among them would become extremely difficult, it was decided to

have three categories of reading demand: low, medium, and high. Furthermore, it was very important to be able to document that the lowest category included items with lower reading demands than the medium category and, in particular, the highest category. Thus, the process of separating the items into the three categories of low, medium, and high reading demands involved several phases, including a review of the literature about the factors influencing reading demand, a holistic evaluation of the items according to selected indicators of reading demand, coding each item according to the reading demand indicators, and validating the holistic item categorizations through discriminant function analysis.

Holistic Evaluation of the Level of Reading Demands in the TIMSS 2011 Fourth Grade Items

As a first step toward holistically rating the TIMSS fourth grade items according to level of reading demand, the TIMSS & PIRLS International Study Center staff conducted a detailed review of the literature concerning dimensions of reading difficulty in the context of evaluating the reading demands presented by the TIMSS 2011 fourth grade mathematics and science items.

Settling on the best set of indicators to capture the reading demand of the TIMSS items was somewhat challenging, because much of the research on factors influencing reading difficulty is based on continuous text of some length, whereas the TIMSS fourth grade items are short. Also, the preponderance of research about reading difficulty in test items highlights how difficulty can be reduced by using clearer, less complicated language; however TIMSS already makes every effort to avoid unnecessary reading and the language used is no more complex than needs to be to frame the question (and responses for multiple-choice items). Additionally, the reading demand indicators used for this study needed to be applicable across the many languages of the TIMSS countries, which imposed further operational and practical constraints. When considering which of the many factors that influence reading difficulty could be used as indicators of reading demand, the following criteria were kept in mind:

- ◆ Appropriateness for the TIMSS fourth grade items;
- ◆ Generalizability across languages;
- ◆ Likelihood of being applied reliably; and
- ◆ Feasibility within resource and time constraints.

After a detailed discussion, staff articulated a set of indicators that appeared most applicable to evaluating the reading demands of the TIMSS fourth grade items. These included:

1. **Number of Words**—The number of words one must read is a basic feature of reading difficulty included in many well-known readability formulae (e.g., Dale-Chall and Flesch-Kincaid). While it was recognized that the number of words varies across languages, it was assumed that the items with more words in English would also have more words in other languages; therefore the relative reading demands across items would be maintained.
2. **Vocabulary**—A unique feature of reading in the mathematics and science context is that there are specialized vocabularies one must know for complete comprehension, but the use of particular vocabulary terms can contribute to reading demand (Adams, 2003; Bernardo, 2005; Justenson & Katz, 1995; Kane, Bryne, & Hater, 1974). Although there is some debate, both mathematics and science generally are regarded as having specialized languages with their own technical vocabularies, including everyday language that has specific meaning when used in the mathematics or science context (e.g., “difference” and “more” in mathematics; and “stay alive” instead of “survive” in science).
3. **Symbolic Language**—Similar to the specialized vocabulary component, understanding symbolic language requires reading skills that are particularly important in the mathematics and science context (Matteson, 2006). These can include numerals (e.g., 3, 5, 40) as well as other symbols and abbreviations (e.g., +, =, cm).
4. **Visual Displays**—The TIMSS achievement items contain a range of visual displays that students need to interact with to varying degrees in order to successfully complete the items. The complexity or density of a visual display impacts reading difficulty (Matteson, 2006; Mosenthal & Kirsch, 1998). Visual displays included the following: 1) pictorial representations of real world things, 2) geometric shapes and figures, 3) models and diagrams, 4) tables, and 5) graphs.

The above indicators of reading difficulty were used to holistically rate the TIMSS items according to their reading demands. Taking the specified components into account, ten members of the TIMSS & PIRLS International Study Center with backgrounds in measurement, reading, mathematics, and science used a holistic approach to evaluate the reading demands required by each item as low, medium, or high. According to the holistic scoring approach, the categorization was based on the overall impression of the reading difficulty of the item, with the proviso to assign about the same number of items to each category to ensure stability in the analyses. After independently rating each of the items, the entire team met to reconcile results and reach group consensus on the holistic rating of each item as low, medium, or high.

Empirical Data About the Reading Difficulty Factors Present in Each Item

The next phase of the study involved validating the holistic ratings, by coding each of the items according to the four dimensions of reading difficulty. The TIMSS & PIRLS International Study Center staff developed a draft coding guide to identify and quantify the difficulty factors present in each of the TIMSS fourth grade mathematics and science items. This draft underwent a series of internal reviews. Then, at the June 2012 meeting of TIMSS and PIRLS 2011 National Research Coordinators (NRCs) in Singapore, NRCs were led through the draft coding guide, and they suggested ways in which the coding guide could be further improved.

Most importantly, there was a thorough debate among the NRCs about the intersection of reading and mathematics, particularly in the areas of symbolic language and geometric shapes. That is, when students are asked simply to solve an equation for “X,” are they reading the language of mathematics or only “doing mathematics” without any reading? Similarly, when asked to analyze the attributes of a triangle, are students engaging in an activity similar to reading a diagram, or is that only “doing mathematics?” Based on the literature review and the discussion, the TIMSS & PIRLS International Study Center decided to code these aspects of reading difficulty with separate codes so that analyses could be done with or without these aspects should researchers be interested.

The major indicators of reading demand in the final coding guide applied to each TIMSS fourth grade item were as follows:

- ◆ The number of words (anywhere in the item, including titles of graphics and labels);
- ◆ The number of different symbols (e.g., numerals, operators);
- ◆ The number of different specialized vocabulary words; and
- ◆ The total number of elements (density) in the visual displays (e.g., diagrams, graphs, tables).

The coding guide was implemented to document the reading demands of each of the TIMSS fourth grade mathematics and science items (see *Technical Appendix A: Quantifying the Reading Demands of the TIMSS 2011 Fourth Grade Mathematics and Science Items*). In order to ensure that all dimensions of difficulty were correctly coded, each item was coded independently by two TIMSS or PIRLS senior staff members, who then reconciled any discrepancies. Exhibits 2.1 and 2.2 provide examples of the detailed coding applied to each item.

Exhibit 2.1: Mathematics Item Coding Example

Mathematics Item Coding Example

If the string in the diagram above is pulled straight, which of these is closest to its length?

(A) 5 cm
(B) 7 cm
(C) 8 cm
(D) 9 cm

Number of Words: 18

Symbolic Language: 11 different symbols (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, cm)

Visual Display: Pictorial Representation with a density of 2 (string, ruler) and a “necessary” level of interaction

Exhibit 2.2: Science Item Coding Example

Science Item Coding Example

Stephanie has a balance and four cubes (1, 2, 3, 4). The cubes are made of different materials.

She puts two cubes at a time on the balance and observes the following results.

What can she conclude about the weight of cube 2?

(A) It is heavier than cubes 1, 3, and 4.
(B) It is heavier than cube 1 but lighter than cubes 3 and 4.
(C) It is heavier than cube 3 but lighter than cubes 1 and 4.
(D) It is heavier than cube 4 but lighter than cubes 1 and 3.

Number of Words: 74

Symbolic Language: 4 different symbols (1, 2, 3, 4)

Vocabulary: 2 different terms (balance, cube)

Visual Display: 3 models, each with a density of 3 (1 balance, 2 cubes) and a “necessary” level of interaction

Discriminant Function Analysis

Finally, in order to validate the holistic categorizations of the items according to low, medium, and high reading demands, the data from coding the reading demands were used to conduct a discriminant function analysis (DFA). Group membership according to the holistic ratings was predicted using the four reading demand indicators: number of words, number of different symbols, number of different specialized vocabulary terms, and number and density of visual displays. Exhibit 2.3 presents the results of the DFAs for mathematics and science. For both subjects, the first discriminant function was sufficient to discriminate between the item groups, and the number of words was the indicator that loaded most heavily on this function.

Exhibit 2.3: Discriminant Function Analysis Results

Mathematics			Science		
Loading of Reading Demand Indicators on Discriminant Functions			Loading of Reading Demand Indicators on Discriminant Functions		
Reading Demand Indicators	Function		Reading Demand Indicators	Function	
	1	2		1	2
Total Number of Words	.897	-.137	Total Number of Words	.889	-.407
Sum of Visual Display Density and Interaction Values	.327	.360	Sum of Visual Display Density and Interaction Values	.489	.807
Number of Unique Symbols	.203	-.376	Number of Unique Symbols	.200	.333
Number of Unique Technical Words	.016	.806	Number of Unique Technical Words	.060	.282

Exhibit 2.4 presents the DFA classification results, which show that the reading demand indicators were effective in recovering the low, medium, and high holistic categorizations. The predicted categories largely matched the holistic categories, with agreement on 82 percent of the items for mathematics and 77 percent of the science items.

Exhibit 2.4: Discriminant Function Analysis Classification Results—Confirmation of Item Classification by Reading Demands Based on Holistic Evaluation and Predicted by Discriminant Function Analysis*

Mathematics		Reading Demand Group Predicted by DFA		
Reading Demand Group— Holistic Evaluation		Low	Medium	High
		Low	53	9
Medium	11	46	2	
High	0	9	45	

Science		Reading Demand Group Predicted by DFA		
Reading Demand Group— Holistic Evaluation		Low	Medium	High
		Low	41	12
Medium	6	56	2	
High	0	18	33	

* Shaded cells show number of holistically evaluated items in agreement with DFA.

Characteristics of Reading Demands in the TIMSS 2011 Fourth Grade Items

As explained earlier, the TIMSS 2011 fourth grade mathematics and science items typically do not have heavy reading demands, although some can be somewhat challenging. Exhibit 2.5 shows the characteristics of the items in terms of the four indicators of reading difficulty used in this study. On average, the total number of words (including all words appearing anywhere in the item) was relatively low, especially for mathematics, which included some items simply asking for computation. The average number of words in the mathematics items was 25 with a maximum of 84 words, and the average for the science items was 41 words with a maximum of 151 words. As would be anticipated, the mathematics items had more symbolic language (e.g., numerals

and operators) than the science items, but still averaged only 5 unique symbols per item. Because the language in these items was intended to be at the fourth grade level or lower, occurrences of specialized mathematics and science terms were low (2 to 3 terms per item on average). Finally, the number of visual displays refers to the total density or number of elements in the visual display or displays in the item. Because the TIMSS fourth grade mathematics framework includes geometric shapes as well as data displays (e.g., tables and graphs), the mathematics items more often included visual displays with a number of elements (8 on average) than the science items (3 on average).

Exhibit 2.5: Indicators of Reading Difficulty for the TIMSS 2011 Fourth Grade Items

Number	Mathematics Items (n=175)		Science Items (n=168)	
	Mean	Maximum	Mean	Maximum
Total Words*	25	84	41	151
Different Symbols**	5	30	1	15
Different Specialized Terms	2	10	3	13
Density Visual Displays***	8	46	3	27

* Includes all words appearing anywhere in the item—stem, question, response categories, and visual displays (e.g., exhibit titles, labels)

** Includes numerals, signs of operations, units (e.g., abbreviations such as cm), variables (e.g., X), and labels (e.g., A for angle A).

*** Includes all elements in the visual display(s).

Exhibits 2.6a through 2.6f present examples of TIMSS 2011 mathematics and science items with low, medium, and high reading demands, respectively. The items in the low reading demands category typically were very straightforward, including mathematics computation items with hardly any words, short constructed response items where the question asked for the answer in a word or phrase, and basic multiple choice items with a question and short options. The items in the medium category had more words, and also often had diagrams or geometric figures especially for mathematics. In comparison, the items in the high category usually had both more words and more complex visual displays.

Exhibit 2.6: Examples of TIMSS Fourth Grade Mathematics and Science Items Categorized as Having Low, Medium, and High Reading Demands

Exhibit 2.6a: *Example TIMSS Fourth Grade Mathematics Item Categorized as Having Low Reading Demands*

M051203

$23 \times 19 =$

Answer: _____

Exhibit 2.6b: *Example TIMSS Fourth Grade Science Item Categorized as Having Low Reading Demands*

S031254

Which of these animals has a young form that looks the most like the adult form?

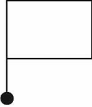
- (A) moth
- (B) human
- (C) frog
- (D) butterfly

SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

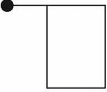
Exhibit 2.6: Examples of TIMSS Fourth Grade Mathematics and Science Items Categorized as Having Low, Medium, and High Reading Demands (Continued)

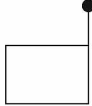
Exhibit 2.6c: Example TIMSS Fourth Grade Mathematics Item Categorized as Having Medium Reading Demands

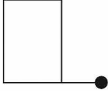
M031071




Which of the following shows the position of the shape above after a half turn or 180° rotation?

(A) 

(B) 

(C) 

(D) 

SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

Exhibit 2.6: Examples of TIMSS Fourth Grade Mathematics and Science Items Categorized as Having Low, Medium, and High Reading Demands (Continued)

Exhibit 2.6d: *Example TIMSS Fourth Grade Science Item Categorized as Having Medium Reading Demands*

S031421

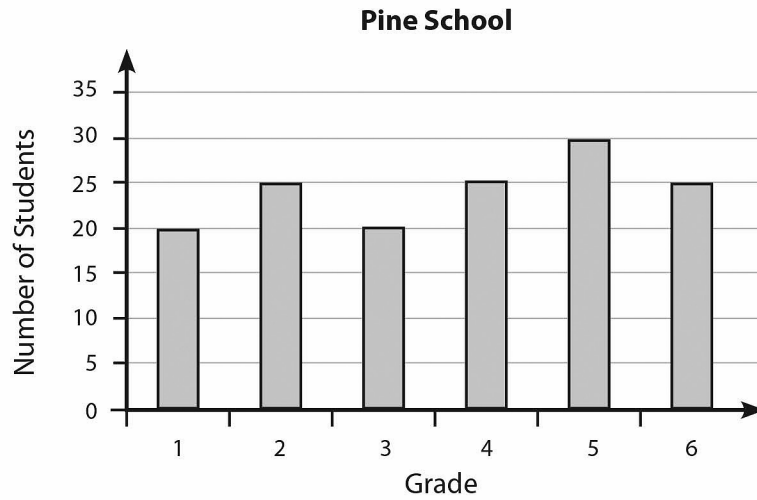
Some of the materials below will burn and some will not.
Put an X in the box next to the materials that will burn.
(You may put an X in more than one box.)

- water
- wood
- sand
- gasoline
- air

SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

Exhibit 2.6e: Example TIMSS Fourth Grade Mathematics Item Categorized as Having High Reading Demands

The graph shows the number of students at each grade in the Pine School.



In the Pine School there is room in each grade for 30 students. How many more students could be in the school?

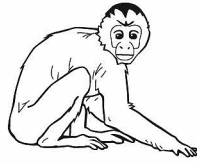
- (A) 20
- (B) 25
- (C) 30
- (D) 35

M051117

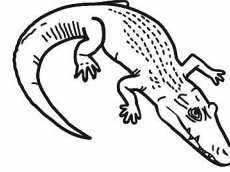
SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

Exhibit 2.6: Examples of TIMSS Fourth Grade Mathematics and Science Items Categorized as Having Low, Medium, and High Reading Demands (Continued)

Exhibit 2.6f: Example TIMSS Fourth Grade Science Item Categorized as Having High Reading Demands



Monkey



Crocodile



Grasshopper



Octopus

Answer the following questions using the animals shown above. Write the name for the correct animal in the spaces below.

Which animal has an internal skeleton and produces milk for its young?

Which animal has an external skeleton and three pairs of legs?

Which animal has a soft body and no skeleton?

S031233

SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

Interaction Between the Levels of Reading Demands and the TIMSS 2011 Content and Cognitive Domains at the Fourth Grade

It should be recognized that the nature of the content and cognitive domains in TIMSS frameworks can heavily influence the reading demands required by the items. Exhibit 2.7 shows the content and cognitive domains for the TIMSS 2011 mathematics and science items at the fourth grade. Of these two dimensions of the mathematics and science assessments, the content domains describe in some detail the major content to be assessed and the cognitive domains describe the thinking skills the students should be using within the content domains. The TIMSS 2011 mathematics and science assessments each encompassed three content domains and three cognitive domains. The cognitive domains were the same for mathematics and science and at the fourth grade and they had the same amount of emphasis.

Exhibit 2.7: TIMSS 2011 Content and Cognitive Domains at the Fourth Grade Percentages Devoted to Each Domain

Mathematics		Science		Mathematics and Science	
Content Domains	Percentage of Assessment	Content Domains	Percentage of Assessment	Cognitive Domains	Percentage of Assessment
Number	50%	Life Science	45%	Knowing	40%
Geometric Shapes and Measures	35%	Physical Science	35%	Applying	40%
Data Display	15%	Earth Science	20%	Reasoning	20%

Exhibit 2.8 presents the distributions by content domain of the TIMSS 2011 mathematics and science fourth grade items categorized as having low, medium, and high reading demands. For mathematics, there is a clear interaction between the content domain and the reading demands of the items. Half of the assessment is devoted to assessing the number domain, with 50 percent of the number items being classified as having low reading demands because many ask only for computation or familiarity with basic number concepts. The geometry and measurement items (about a third of the assessment) often were categorized as medium (49%), because the framework calls for assessing a variety of understandings related to points, lines, and angles as well as two- and

three-dimensional shapes. Therefore, the typical item in this content domain includes a figure with an associated question, with some items being very straightforward and others more complicated, but primarily the items were in-between. Finally, although only a small part of the assessment is devoted to data display, most of these items (85%) were categorized as having high reading demands because, consistent with the framework topics, these items typically involved reading and interpreting data from relatively dense visual displays, including tables, pictographs, bar graphs, and pie charts.

Exhibit 2.8: TIMSS Fourth Grade Mathematics and Science Items by Content Domains and Low, Medium, and High Reading Demands

Reading Demands	Mathematics Content Domains				Science Content Domains			
	Number	Geometric Shapes and Measures	Data Display	Total Math	Life Science	Physical Science	Earth Science	Total Science
Low	50%	28%	4%	35%	36%	26%	30%	32%
Medium	30%	49%	12%	34%	35%	33%	55%	38%
High	20%	23%	85%	31%	28%	41%	15%	30%
Total	88	61	26	175	74	61	33	168

Interestingly, the degree of difficulty of the TIMSS mathematics items across content domains may be unexpected in light of the relative levels of reading demands. As shown in the report containing the *TIMSS 2011 International Results in Mathematics* (Mullis, Martin, Foy, & Arora, 2012), the average percent correct across the mathematics items overall was 50 percent, with 47 percent correct, on average, for the number items, 49 percent for the geometric shapes and measures items, and 58 percent for the data display items. On other hand, the degree of reading demands in the mathematics items by content domains has some relationship with the emphasis on the topics in these three content domains in the curricula across countries. According to data published in the *TIMSS 2011 International Results in Mathematics*, across countries, National Research Coordinators reported that of the eight number topics, on average, six (75%) were included in the curriculum; of the seven geometry topics, five (71%) were included, and of the three data display topics, two (67%) were included. According to their teachers, the percentage of

students that had been taught the TIMSS topics, averaged across each content domain, was lower for geometric shapes and measures (65%) than for both number and for data display (76%). The NRC and teacher data indicate less curricular emphasis on the TIMSS topics in geometric shapes and measures than in number, whereas the TIMSS items in geometric shapes and measures are more likely than the number items to have medium rather than low reading demands. The situation with data display is more difficult to interpret because it receives a small emphasis in the assessment (15%) and only has three topics. Although the three data display topics seem to be present in the curricula and classrooms of the TIMSS countries, this probably represents only a small part of students' instruction in mathematics and the TIMSS items in this content domain are likely to have relatively high reading demands.

Looking across the science content domains, the life science items—covering topics about the characteristics, processes, and cycles of living things and comprising nearly half of the assessment (45%)—were relatively well distributed according to reading demand, although with a tendency toward lower or medium rather than high levels of reading difficulty. The opposite was found for the physical science items (35% of the assessment) which often involve physical phenomena that can be presented via models or diagrams. Thus, items in the physical science content domain were more likely to have medium and high reading demands. Finally, the one-fifth of the assessment devoted to earth science was well-balanced with most items categorized as having medium reading demands, but some with low and some with high reading demands.

For science at the fourth grade, as reported in *TIMSS 2011 International Results in Science* (Martin, Mullis, Foy, & Stanco, 2012) the average percent correct across countries was 48 percent overall, and very similar across content domains—life science (48%), earth science (46%), and physical science (49%). However, the content areas more likely to have TIMSS items with high reading demands were the content areas emphasized least in the curricula across countries. The life science items were most likely to have low reading demands, earth science items medium reading demands, and physical science items high reading demands. This corresponds with the curricular emphasis on the science content areas, with life science being emphasized more than earth science, and earth science, in turn, emphasized more the physical science. As presented in the *TIMSS 2011 International Results in Science*, across countries, National Research Coordinators reported, on average, that of the six life science topics, five (84%) were included in the curriculum for all students; of the six earth science topics,

four (67%) were included; and of the eight physical science topics, five (63%) were included. Consistent with the country reports, according to their teachers, on average, the percentage of students taught the TIMSS topics was highest for life science (75%), next highest for earth science (63%), and lowest for physical science (57%).

Exhibit 2.9 presents the distributions by cognitive domain of the TIMSS 2011 mathematics and science fourth grade items categorized as having low, medium, and high reading demands. In general, the patterns are similar for mathematics and science. The two-fifths of the items measuring the knowing domain (e.g., recall, recognize, compute, classify/order) were more likely to have low reading demands (61% in mathematics and 56% in science); the two-fifths of the items measuring the applying domain (e.g., represent, model, and solve standard problems) were more likely to have medium reading demands; and the one-fifth of the items measuring reasoning (e.g., analyze, synthesize, justify, and solve problems in unfamiliar or complex contexts) were most likely to have high reading demands (59% in mathematics and 76% in science). Items measuring students’ ability to apply their knowledge of content and procedures and, especially those requiring reasoning, need to include some information in the form of words or visual displays as to the problems situation. The items measuring reasoning often were based on scenarios or situations reflecting school or daily experiences.

Exhibit 2.9: Percentage of TIMSS Fourth Grade Mathematics and Science Items by Cognitive Domains and Low, Medium, and High Reading Demands

Reading Demands	Mathematics Cognitive Domains				Science Cognitive Domains			
	Knowing	Applying	Reasoning	Total Math	Knowing	Applying	Reasoning	Total Science
Low	61%	24%	6%	35%	56%	18%	7%	32%
Medium	21%	45%	35%	34%	35%	33%	17%	38%
High	17%	31%	59%	31%	28%	41%	76%	30%
Total	70	71	34	175	74	61	29	168

For the cognitive domains in mathematics and science, the items were more difficult across the cognitive areas from knowing, to applying, and then reasoning. For mathematics, the average percent correct across countries was

55 percent for the items classified in the knowing domain, 50 percent for items classified in the applying domain, and substantially lower for items in the reasoning domain—40 percent (Mullis, Martin, Foy, & Arora, 2012). In science, the average percent correct across countries was 53 percent for knowing, 46 percent for applying, and 41 percent for reasoning (Martin, Mullis, Foy, & Stanco, 2012). Certainly, it can be considered that both the complexity of the cognitive tasks and the increase in reading demands both contribute to the substantial difficulty of the reasoning items.

Generalizability Across Countries of the Item Categorizations According Low, Medium, and High Reading Demands

Because the study of the impact of item reading demands on TIMSS fourth grade mathematics and science achievement was conducted using the English language version of the items, the question arises about the generalizability of the results to languages other than English. Clearly, the four indicators of reading demands would not be expected to have identical values in all languages. For example, it is well known that after translation the number and length of words in the TIMSS items varies across languages. However, it is likely that the items with the highest reading demands in English also would have the highest reading demands in other languages. Thus, National Research Coordinators of countries that conducted TIMSS 2011 in languages other than English participated in a Reading Demands Matching Analysis (RDMA) which involved categorizing the items in their languages into three categories of reading demand (highest, medium, and lowest). NRCs were instructed to categorize the items in their languages using a holistic rating process based on the same four indicators of reading demand used in the study (number of words, number of different symbols, number of different specialized vocabulary terms, and number/density of visual displays). Seventeen countries and one benchmarking participant (representing 16 different languages in total) submitted their RDMA categorizations to the TIMSS & PIRLS International Study Center. On average, NRCs reported involving four different raters in the RDMA process. The rating teams included members of their TIMSS or PIRLS teams and/or content area experts, including researchers, curriculum experts, and teachers. A number of countries reported that they valued participating in the process and appreciated viewing the TIMSS items from a different perspective.

The categorizations submitted by the NRCs were compared to the categorizations developed by the TIMSS & PIRLS International Study Center, and there was a very high degree of agreement for both the mathematics and science items. On average, there was 71 percent exact agreement and 98 percent adjacent agreement across countries. Although there were some items (15%) where the level of exact agreement was below 50 percent, these were often items with complicated graphics.

The Impact of Reading Ability on TIMSS Achievement for Items with Low, Medium, and High Reading Demand

Keeping in mind that the level of reading demands in the TIMSS fourth grade mathematics and science items interacts with many other factors, including the difficulty and curriculum coverage of the topics in the content domains as well as the variation in difficulty across the cognitive domains, it is still very interesting to look at performance on the TIMSS mathematics and science items for students of different reading ability and how this relates to the level of reading demands in the items from low to medium to high.

The relationship between level of reading ability and TIMSS mathematics and science achievement by level of reading demand in the items was examined by computing the average percent correct¹ for items in each of the three categorizations of reading demands (low, medium, and high) for students with three levels of reading ability. More specifically, separately for mathematics and for science, the TIMSS & PIRLS International Study Center computed the average percent correct for each group of the fourth grade items classified according to low, medium, and high reading demands for each of three levels of students' reading ability (determined by the lower, middle, and upper terciles of reading achievement on PIRLS 2011 in each country).

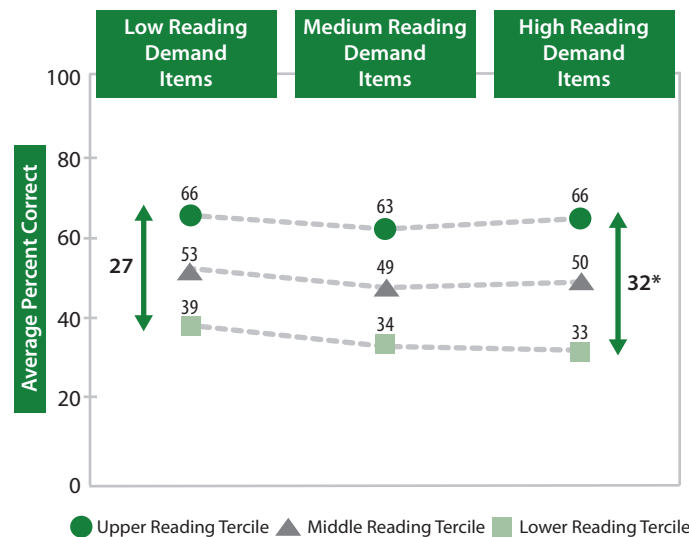
It was expected that the best readers would be unaffected by the reading demands of the items, and therefore would perform similarly on the low, medium, and high demand items, whereas the poorest readers would perform relatively better on low demand items, and less well on high demand items. Some support for these hypotheses was found in the overall and country-by-country results, particularly in mathematics. However, the results varied considerably from country to country and even between mathematics and science within countries.

1 The analyses also were conducted based on achievement scales created for each of the three groups of items using the same scaling approach as used for the TIMSS 2011 content and cognitive domains (i.e., a multi-dimension estimation of performance on high, medium, and low reading demand items using the item parameters from the TIMSS overall concurrent calibration). However, the average percent correct approach used in this paper seemed to provide more easily interpretable results.

Exhibit 2.10 presents line graphs of the international averages across the country-by-country results for TIMSS fourth grade the mathematics items, and Exhibit 2.11 presents the same information for the science items. As would be expected, Exhibit 2.10 for mathematics shows that across the three levels of reading demands (low, medium, and high), the students in the upper reading tercile (top one-third of readers averaged across countries, indicated by **circles**) had higher average mathematics achievement than those in the middle reading tercile (indicated by **triangles**). In turn, the students in the middle reading tercile had higher average mathematics achievement at each level of item reading demand than did those in the lower reading tercile (indicated by **squares**).

Exhibit 2.10: Mathematics Achievement Averaged Across Countries—Fourth Grade

Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands



Results for each tercile averaged across countries.
 * The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands.
 Because of rounding, some results may appear inconsistent.

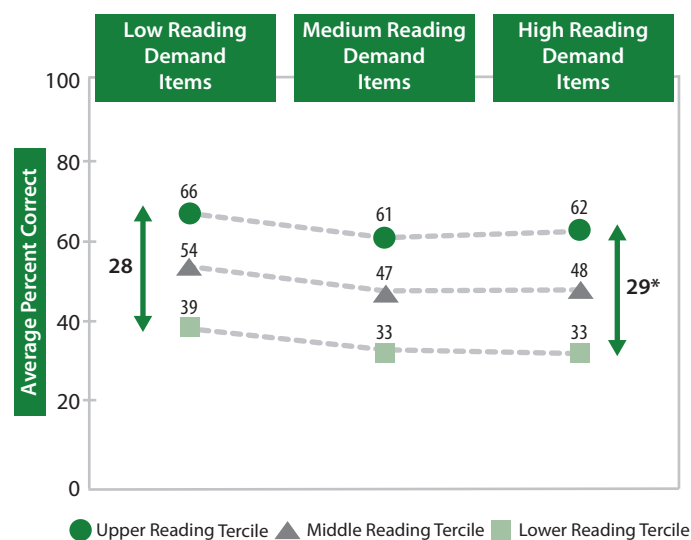
As hypothesized, the average mathematics achievement of the best readers did not vary much by level of reading demands (66% correct on the low demand items, 63% on the medium items, and 66% on the high items). In comparison, the average mathematics achievement of the least proficient readers was 39 percent correct on the items with low reading demands, but lower on the items of medium (34%) and high reading demands (33%). The difference in average achievement between poor and good readers on the low reading

demand items was 27 percentage points. However, this difference increased to 32 percentage points on the high reading demand items. While the poorest readers consistently achieved at a lower level in mathematics than the best readers, they were additionally disadvantaged on the mathematics items that required more reading.

Exhibit 2.11 for science shows results for the poorest readers that are nearly identical to those shown for mathematics, with averages of 39 percent correct on the low demand items and 33 percent correct on both the medium and high demand items. Also, the achievement gaps between the three terciles of readers on the science items with low reading demands are nearly identical to those for

Exhibit 2.11: Science Achievement Averaged Across Countries—Fourth Grade

Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands



Results for each tercile averaged across countries.

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands.

Because of rounding, some results may appear inconsistent.

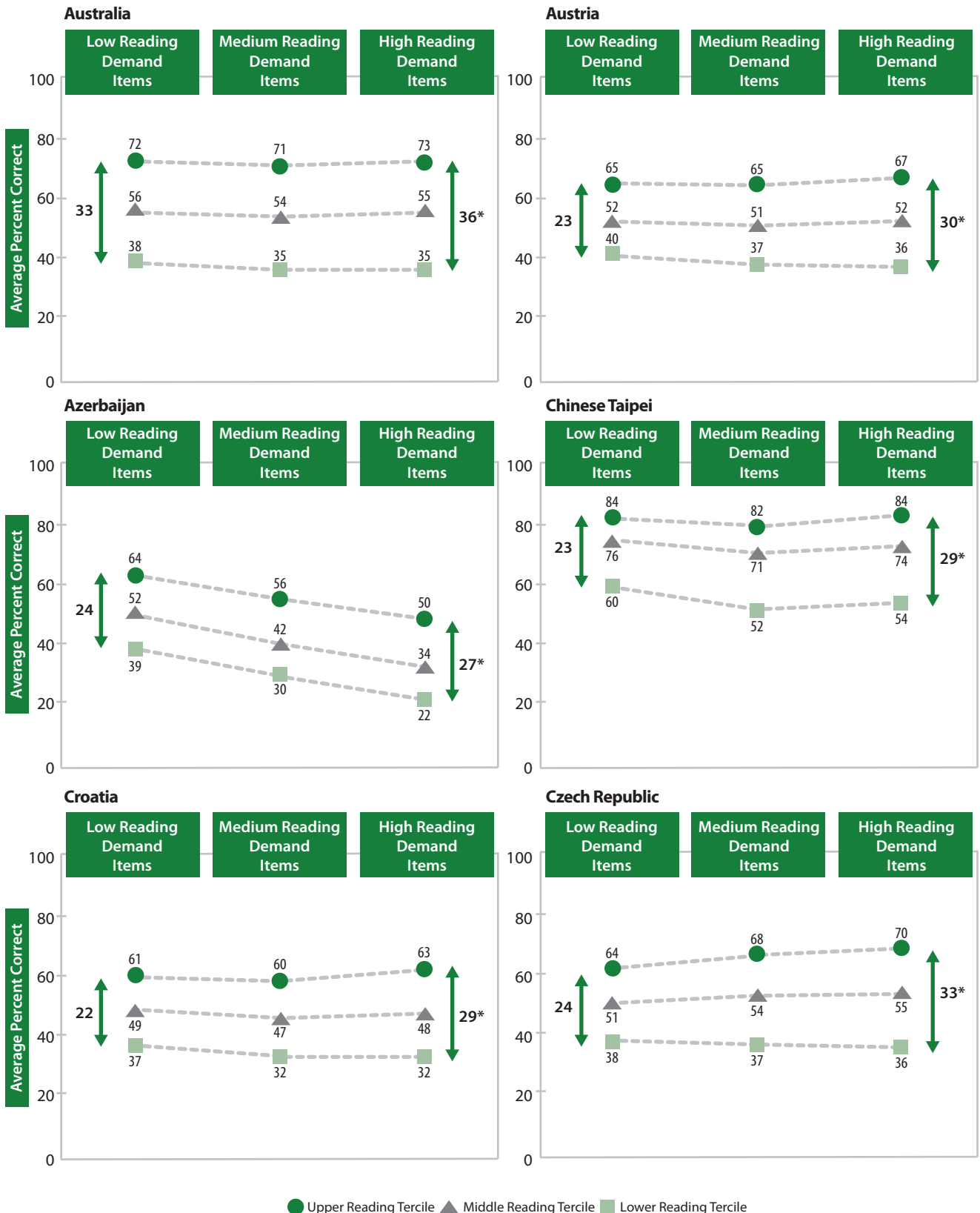
mathematics. However, for the science items, similar to the poorest tercile of readers, the upper and middle terciles of readers also had lower achievement on the medium and high reading demand items. Interestingly for science, there was little difference in the results between the medium and high reading demand items, which may reflect a small difference in reading demands between the medium and high demand items. Because all three terciles of readers were similarly disadvantaged by more reading demands, the gaps in achievement

between the upper tercile of readers compared to the lower tercile of readers was similar at all three levels of reading demand (28% on the low demand and medium demand items, and 29% on the high demand items).

Exhibit 2.12 presents the TIMSS fourth grade mathematics achievement results by level of reading demand and tercile of PIRLS reading achievement for each of the 34 countries and 3 benchmarking entities that assessed the same fourth grade students with TIMSS and PIRLS in 2011. Looking across countries, in most instances there is a significant difference in average percent correct between the upper and lower reading achievement terciles on the low reading demand items (left side of graphs) and the high reading demand items (right side of graphs) with the difference being larger on the high reading demand items. That is, for most countries, better readers have a significantly greater advantage over poorer readers on mathematics items with high reading demands.

The significant difference in the achievement gap between low and high reading demand items seemed to arise from the expected pattern—the best readers having similar mathematics achievement across all items regardless of level of reading demands, but poor readers performing less well on items with more reading demands—most noticeably in Australia, Austria, Chinese Taipei, Croatia, Hungary, Lithuania, Portugal, the Russian Federation, Saudi Arabia, and the Slovak Republic.

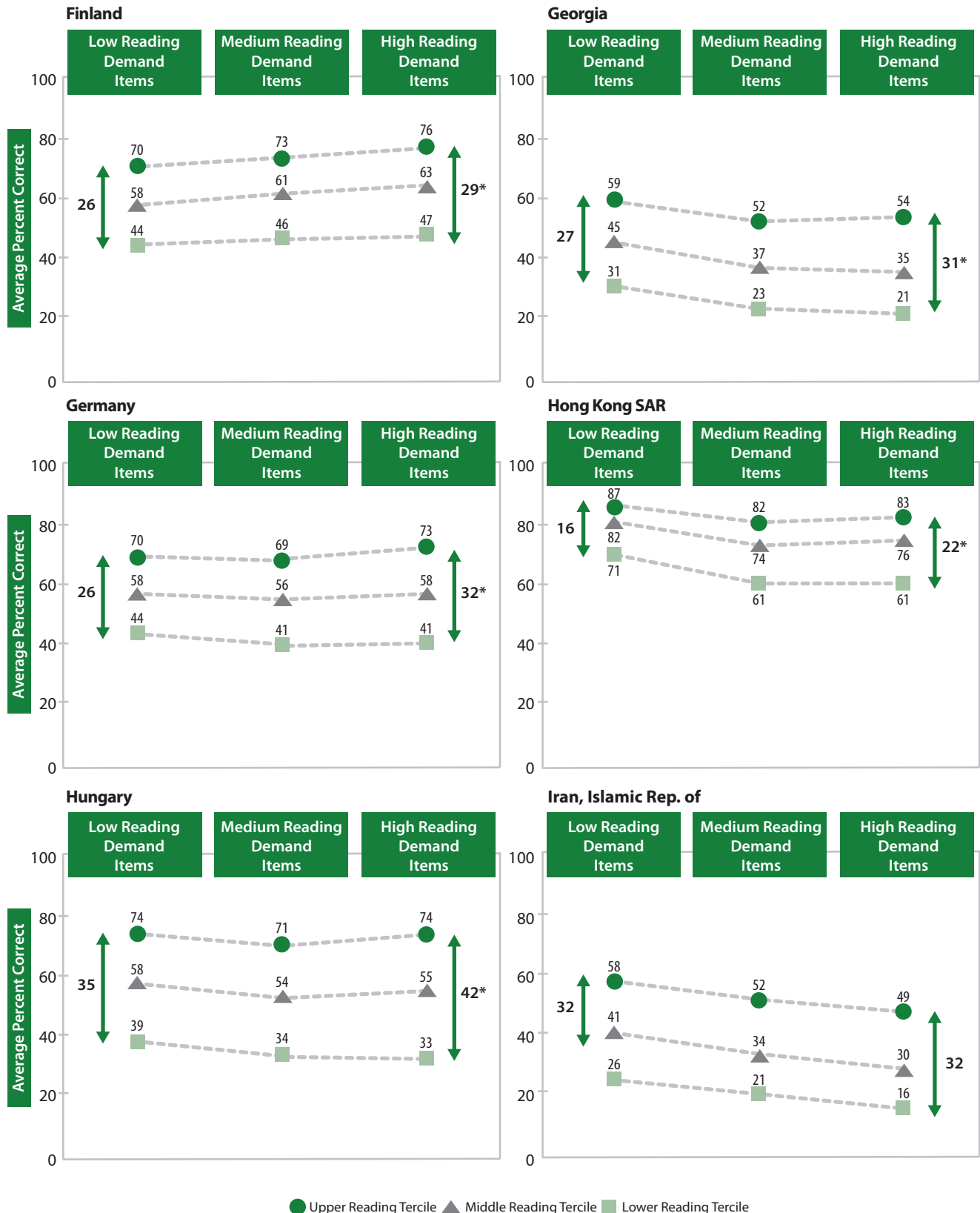
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

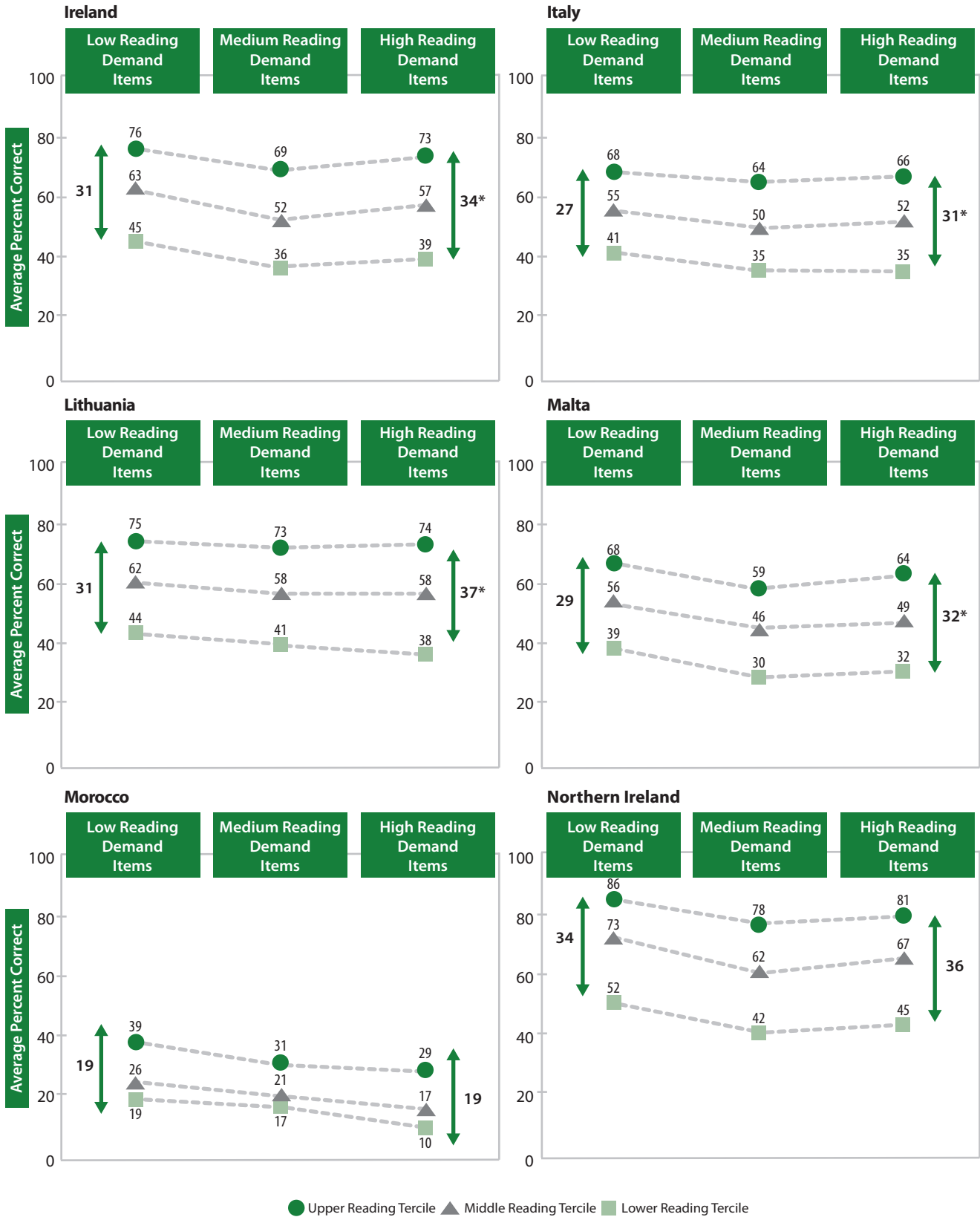
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

*The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

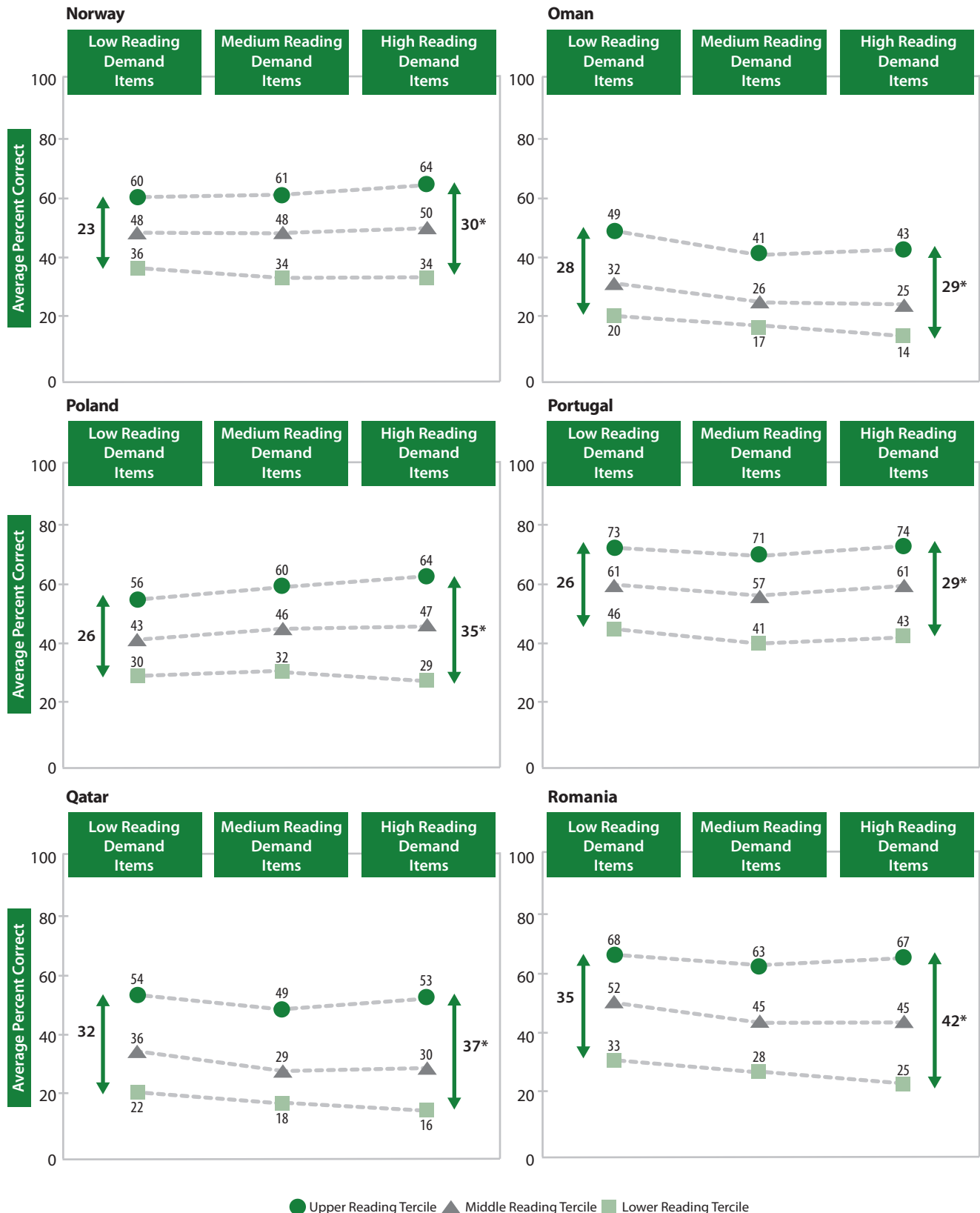
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

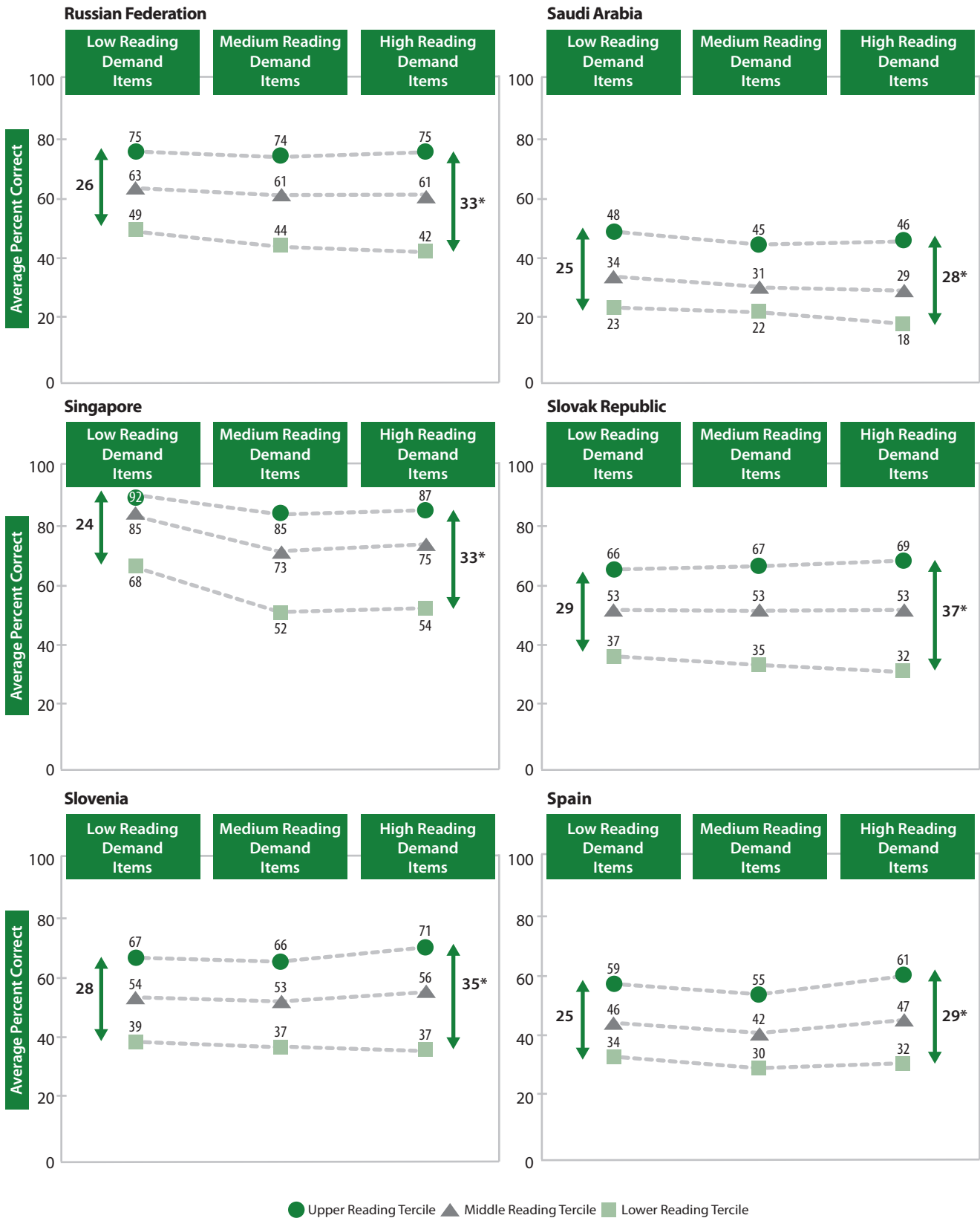
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

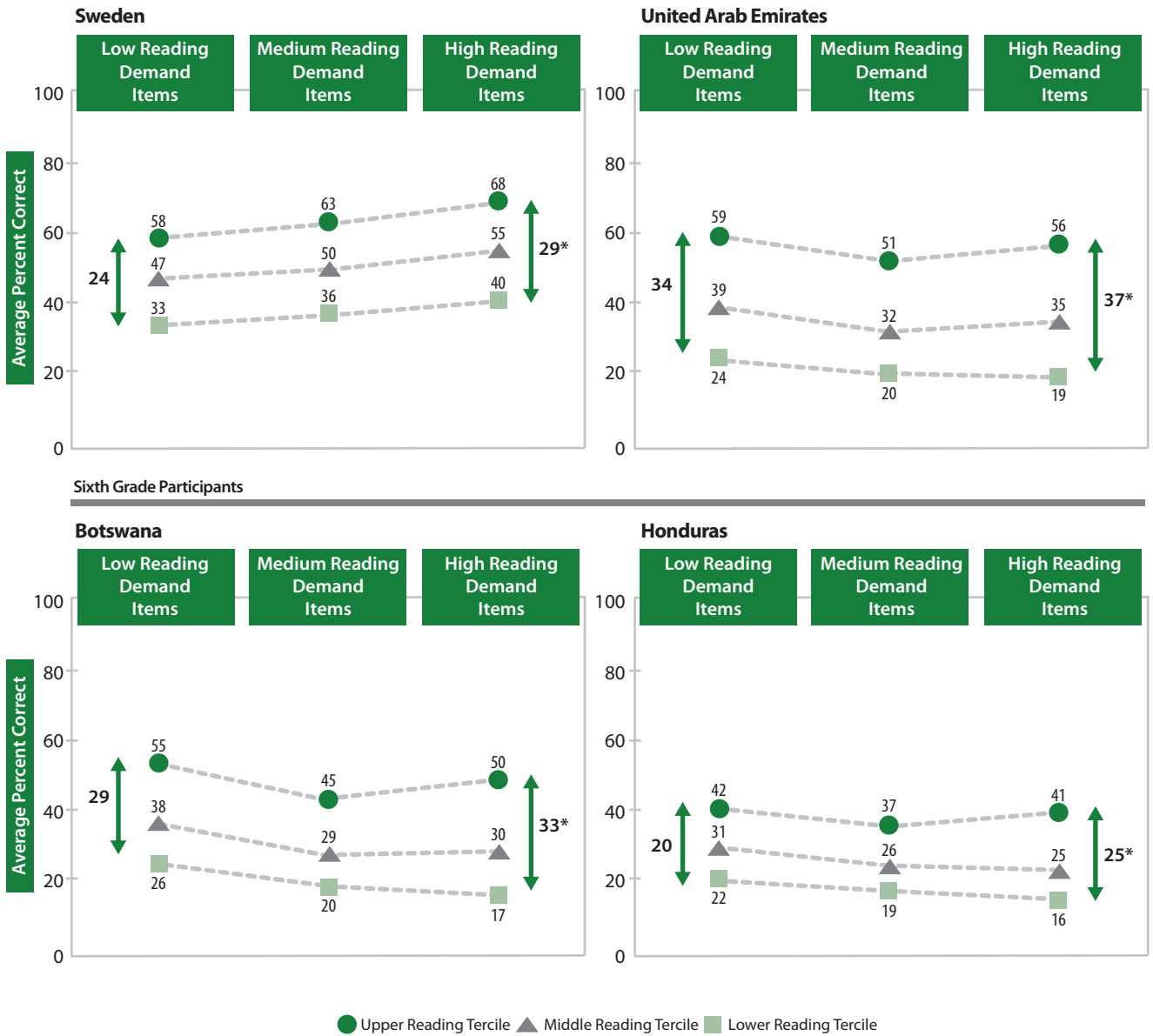
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands

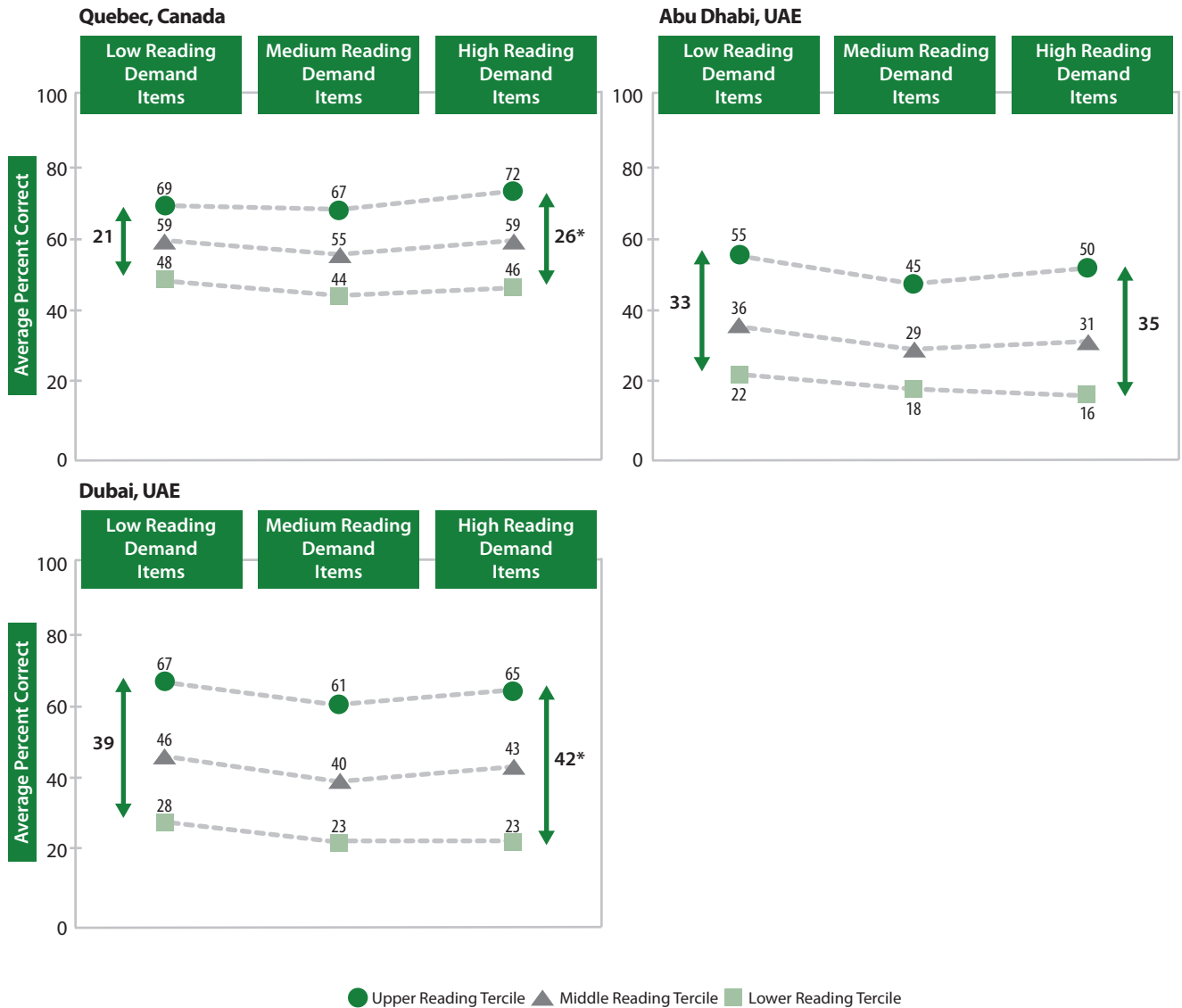


SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Mathematics Items Grouped by Three Levels of Reading Demands

Benchmarking Participants

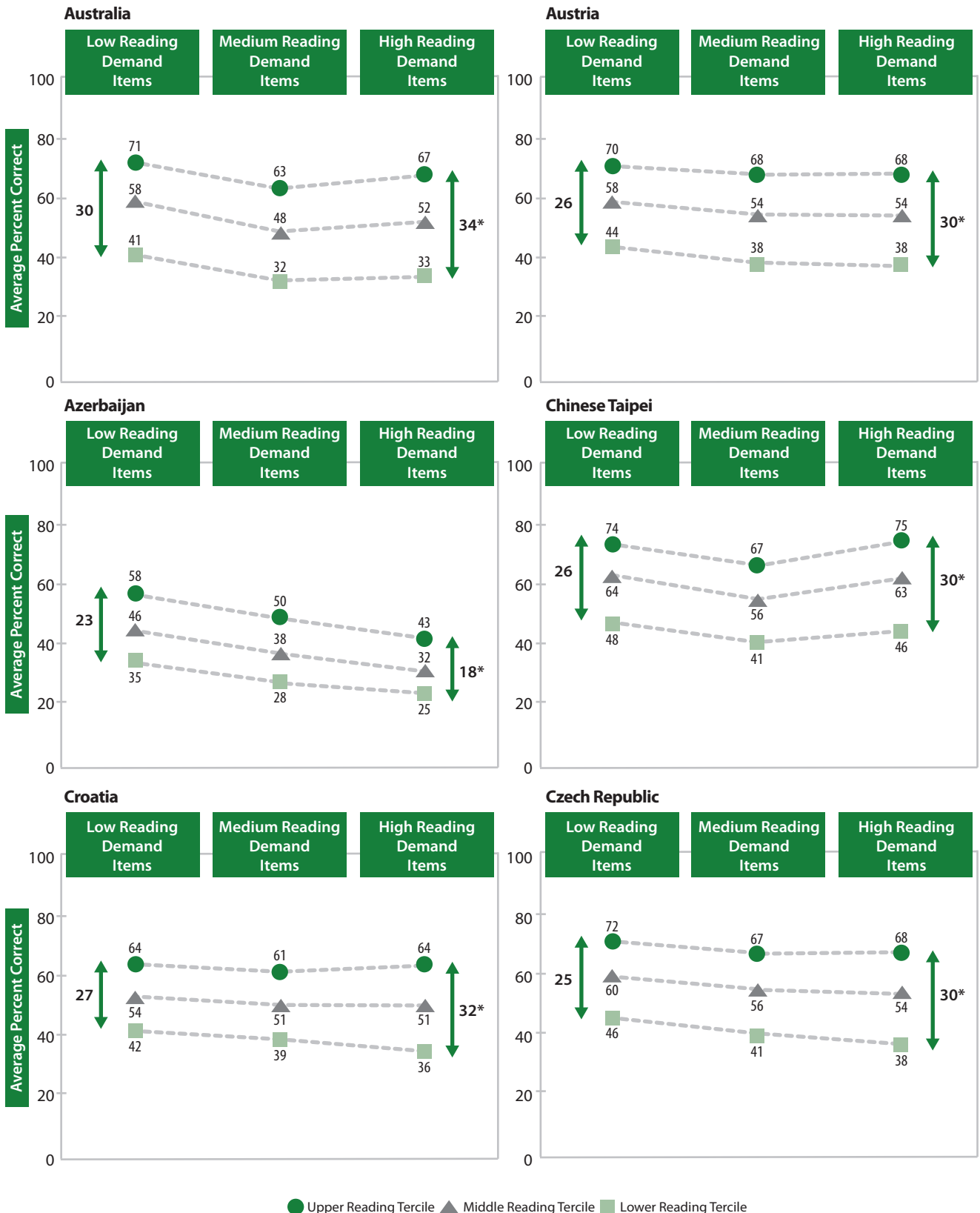


* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

Exhibit 2.13 presents the TIMSS fourth grade science achievement results by level of reading demand and tercile of PIRLS reading achievement for each of the 34 countries and 3 benchmarking entities that assessed the same fourth grade students with TIMSS and PIRLS in 2011. Again, the results may reflect the restricted range in the level of reading demands in the fourth grade science items, or that at the fourth grade science content and reading are very closely linked. Consistent with results averaged across countries (Exhibit 2.11), there were fewer significant differences than in mathematics between the achievement gap for low reading demand items and high reading demand items that show an advantage for better readers over poorer readers on the high reading demand items. However, such significant differences were found in more than half of the countries, including Croatia, the Czech Republic, Hungary, Lithuania, Poland, Romania, the Russian Federation, the Slovak Republic, and Slovenia where the difference was 5 or more percentage points. Interestingly, compared to mathematics, there were more countries, such as Azerbaijan, Georgia, Iran, Morocco, Norway, Romania, and Saudi Arabia, where increased reading demands in the science items was systematically associated with lower achievement for the three levels of readers. In addition, there were a number of countries where performance dropped by about the same amount on the medium and high reading demand items for all three levels of readers.

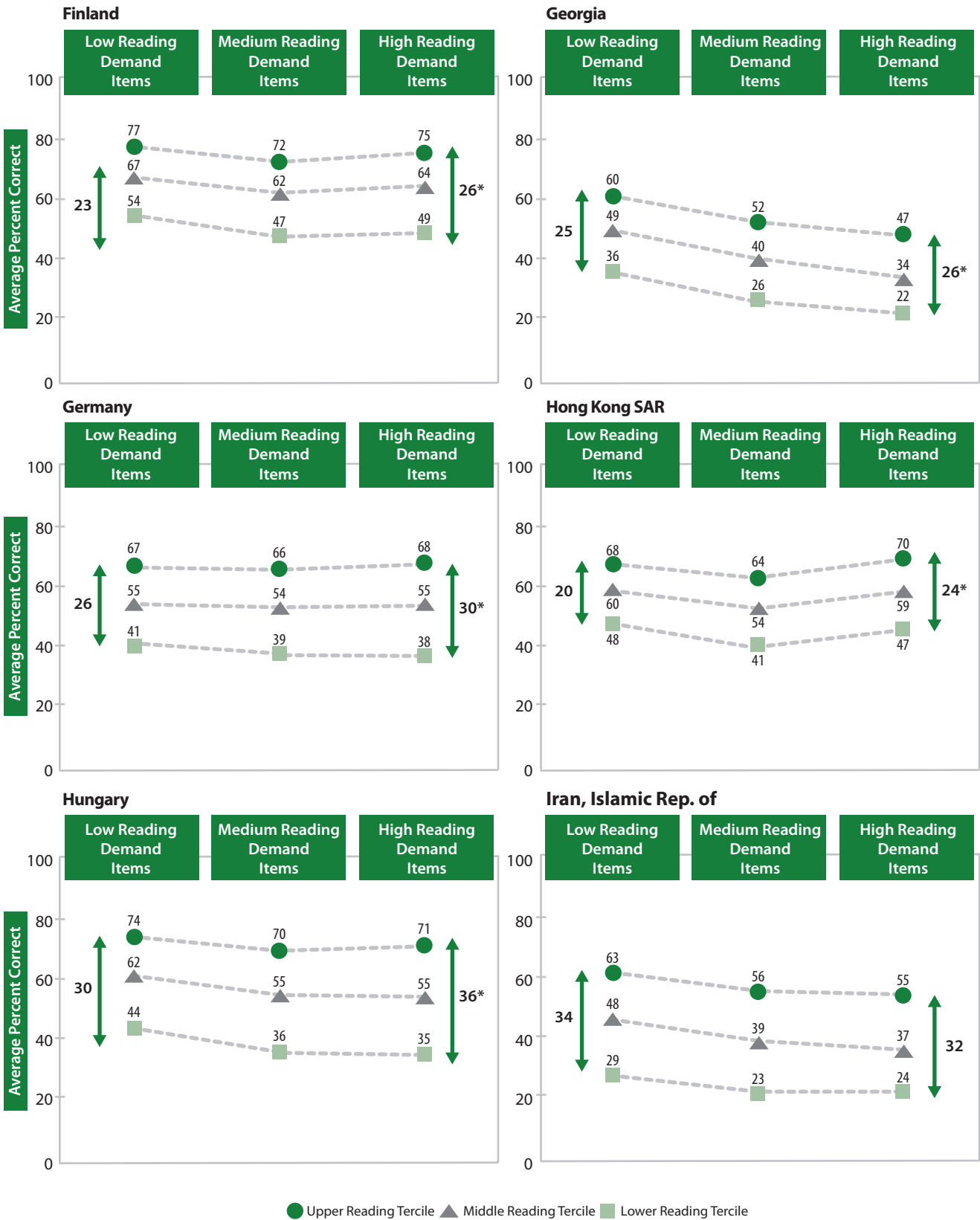
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

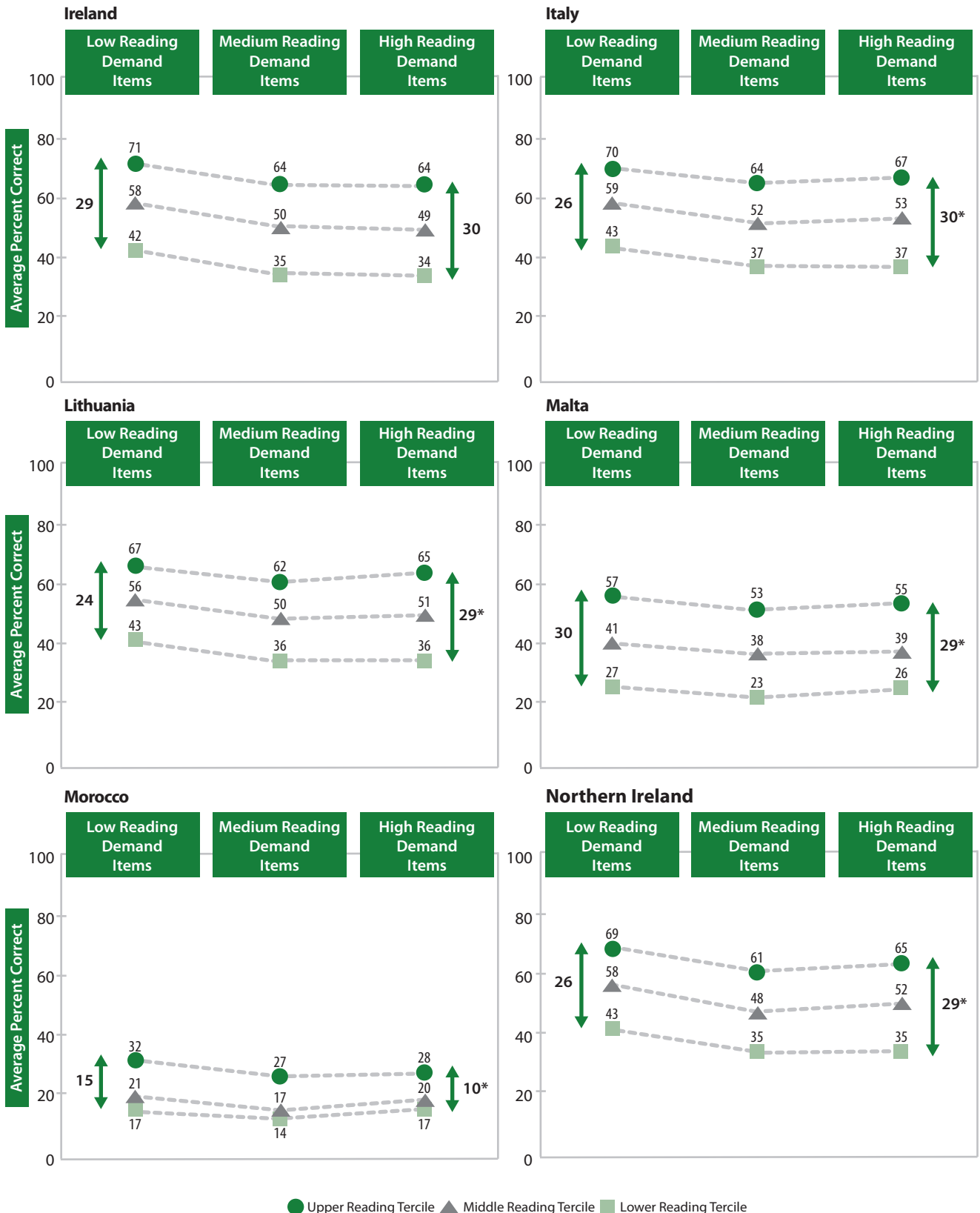
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tertile difference for High Reading Demands is significantly different from the inter-tertile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

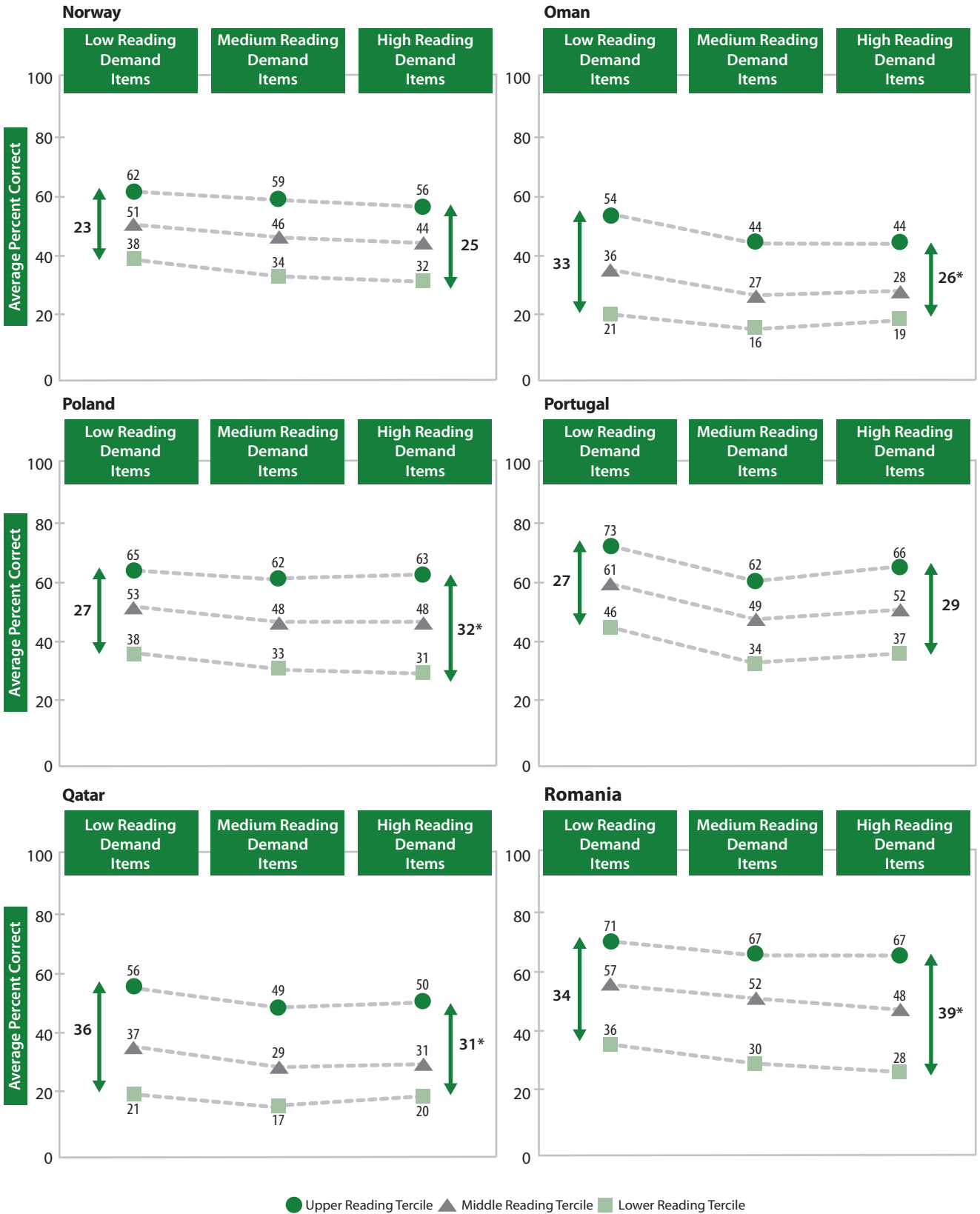
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

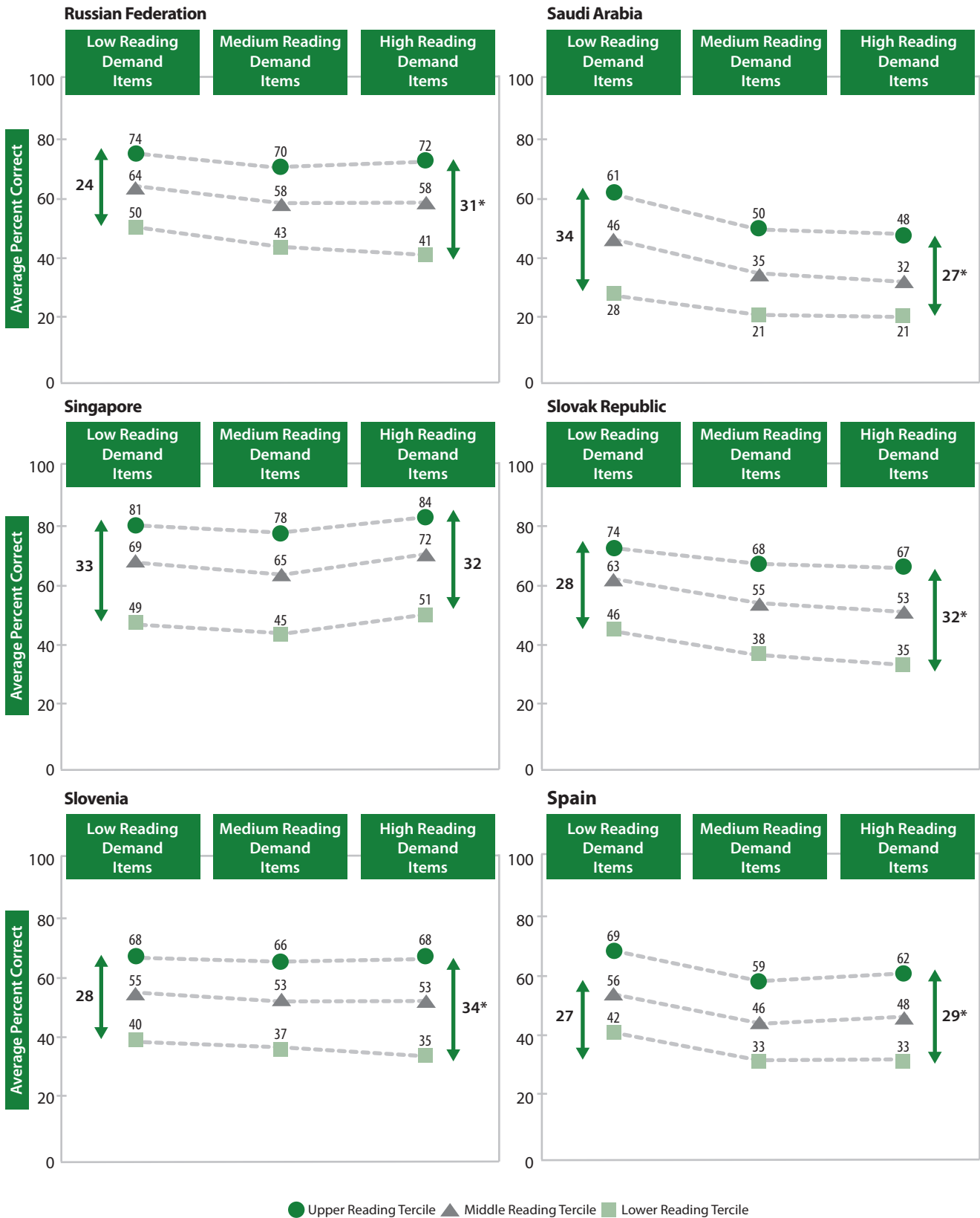
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

*The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

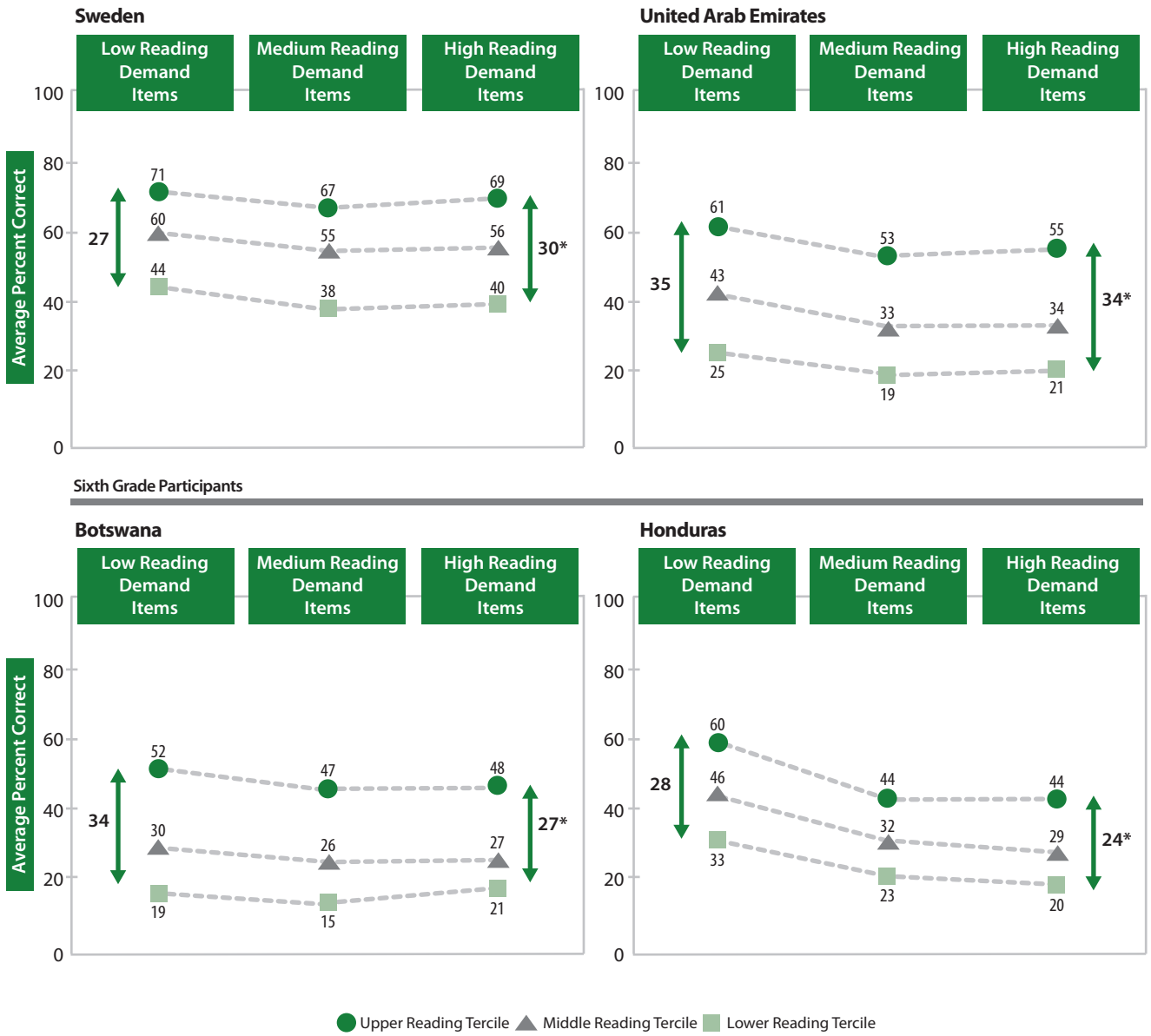
Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands



SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands

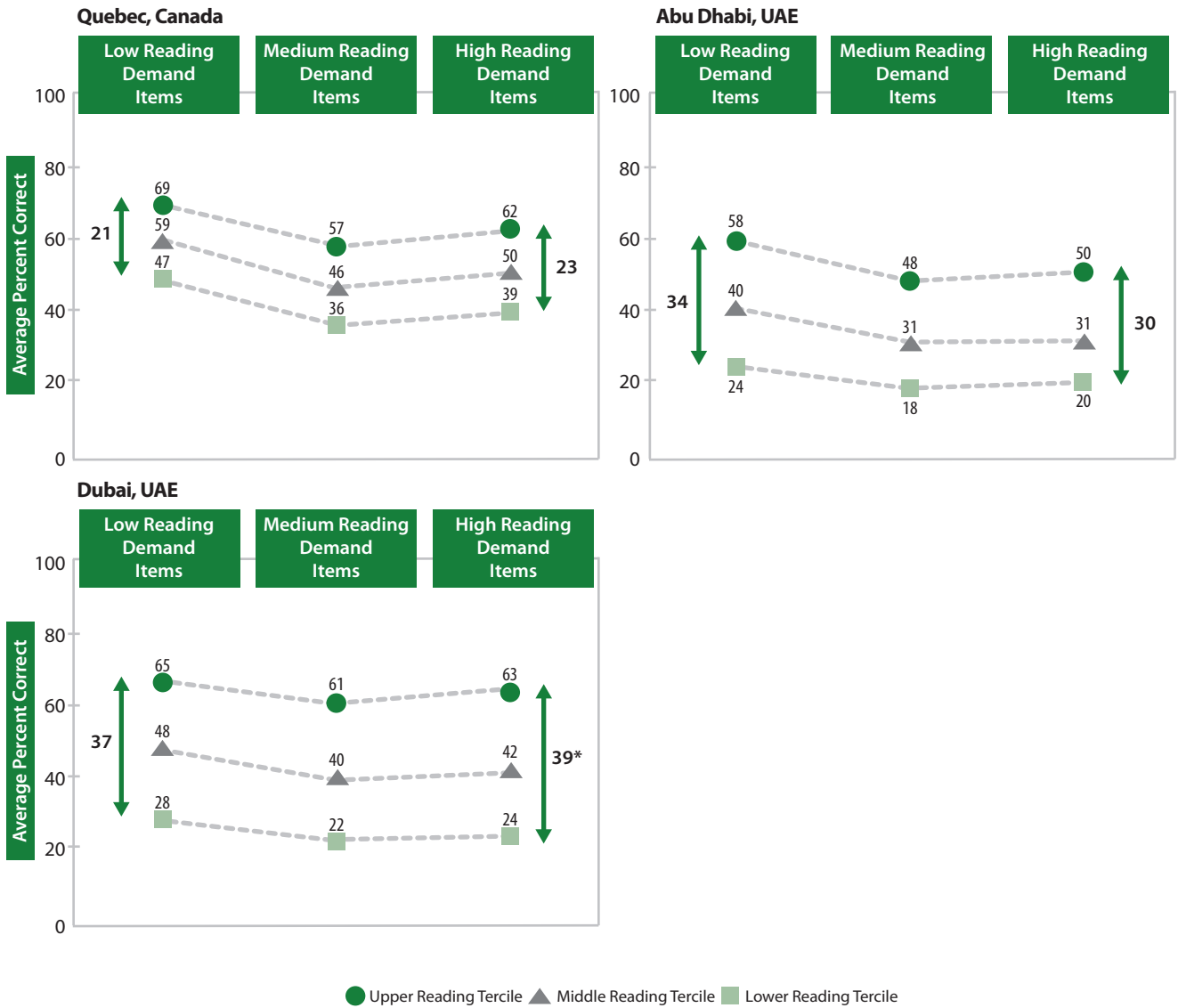


SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Reading Literacy Study – TIMSS and PIRLS 2011

* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

Average Percent Correct for Students at Three Levels of PIRLS Reading Ability on Science Items Grouped by Three Levels of Reading Demands

Benchmarking Participants



* The inter-tercile difference for High Reading Demands is significantly different from the inter-tercile difference for Low Reading Demands. Because of rounding, some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study and Progress in International Literacy Study – TIMSS and PIRLS 2011

Considering the Results

This study hypothesized that the best readers would be unaffected by the reading demands of the items and therefore would perform similarly on the low, medium, and high demand items, whereas the poorest readers would perform relatively better on low demand items, and less well on high demand items. Some support for these hypotheses was found in the overall and country-by-country results, particularly in mathematics.

On average across countries, the mathematics achievement difference between poor and good readers was larger on the high reading demand items than on the low reading demand items. The average mathematics achievement of the best readers did not vary much by level of reading demands, whereas the average mathematics achievement of the least proficient readers was higher on the items with low reading demands than on the items with medium and high reading demands. While the poorest readers consistently achieved at a lower level in mathematics than the best readers, they were additionally disadvantaged on the mathematics items that required more reading.

Also, looking at the results for mathematics country by country, in nearly all instances the difference in average achievement between poor and good readers was larger on the high reading demand items than on the low reading demand items. That is, for most countries, better readers have a significantly greater advantage over poorer readers on mathematics items with high reading demands.

On average across countries in science, all three terciles of readers were similarly disadvantaged by more reading demands, so the gaps in achievement between the upper tercile of readers compared to the lower tercile of readers was similar at all three levels of reading demands. In addition, although achievement was higher on the items with low reading demands, there was little difference in the results between the medium and high reading demand items.

The country-by-country results in science showed that the difference in average achievement between poor and good readers was larger on the high reading demand items than on the low reading demand items in more than half of the countries. However, the difference also was approximately similar in a number of countries.

In summary, much was learned from this research regarding the challenges of educational measurement. Perhaps as a by-product, though still important, was developing procedures to look at the characteristics of the TIMSS fourth grade items through the lens of reading difficulty. In addition to the already

lengthy checklists used to review items for various aspects of content validity and clarity, it is important to scrutinize the TIMSS items from the perspective of “mathematics reading” or “science reading” (e.g., the number of words, the number of different symbols, the load of the technical vocabulary, and the roles that are being played by visual displays).

More important, however, was gaining a deeper understanding of the extreme complexity of the educational endeavor and all of its interconnected parts. Students all over the world are learning mathematics and science, but they are learning these subjects in different ways. In particular, as this research has highlighted, curricular and instructional differences experienced by students can impact item difficulty. As we know, countries’ mathematics and science curricula vary considerably, and the different amounts of emphasis placed on the topics covered has a powerful influence on student learning as well as on student achievement on TIMSS. Beyond that, instructional differences among countries not only affect achievement in the content domains, but also in the cognitive domains. There is an interrelationship between cognitive domain and reading demands, insofar as assessing in-depth content understanding and increased cognitive complexity generally involves greater reading demands; thus, TIMSS mathematics and science items become more difficult for a variety of reasons.

Especially relevant to educational research and policy broadly, though, is that reading is fundamental to further learning. It makes good sense that students who are better readers are therefore better positioned to learn more in mathematics and science as well as in their other subjects. Although the results of this study varied considerably from country to country, and even between mathematics and science within countries, the study showed reading ability to be associated with mathematics and science achievement to an extent that provides support for the idea that greater reading demands can make the fourth grade TIMSS items more challenging for weaker readers. Extrapolating this idea into the broader educational arena raises the question of how much reading intervention might influence learning across the curriculum.

References

- Adams, T.L. (2003). Reading mathematics: More than words can say. *The Reading Teacher*, 56(8), 786-795.
- Bernardo, A.B.I. (2005) Language and modeling word problems in mathematics among bilinguals. *The Journal of Psychology*, 139 (5), 413-425.
- Justenson, J.S. & Katz, S.M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1, 9-27.
- Kane, R.B., Byrne, M.A., & Hater, M.A. (1974). *Helping children read mathematics*. NY: American Book Company.
- Martin, M.O., Mullis, I.V. S., Foy, P., & Stanco, G.M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Matteson, S.M. (2006). Mathematical literacy and standardized mathematical assessments. *Reading Psychology*, 27(2-3), 205-233.
- Mosenthal, P.B. & Kirsch, I.S. (1998). A new measure for assessing document complexity: The PMOSE/IKIRSCH document readability formula. *Journal of Adolescent & Adult Literacy*, 41(8), 638-657.
- Mullis, I.V. S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

