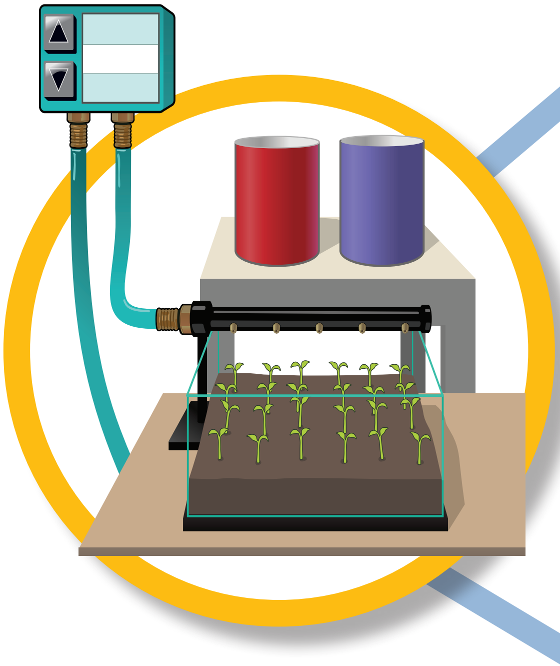
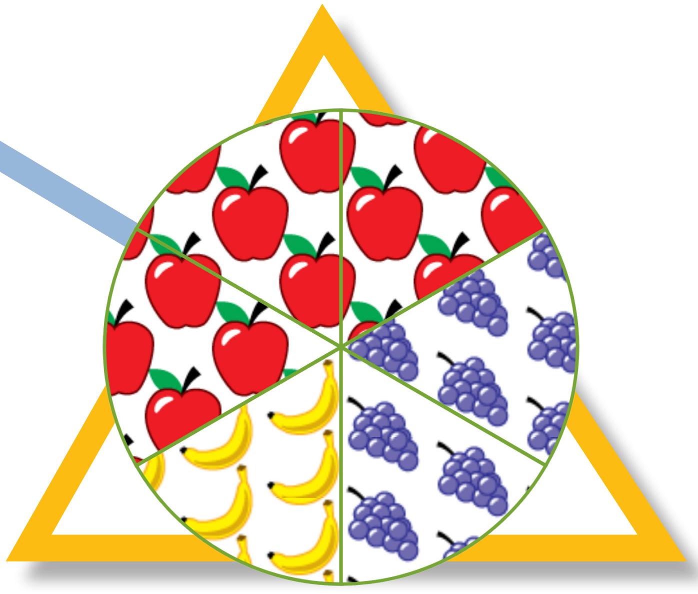
IEA TIMSS 2019



Findings from the TIMSS 2019 Problem Solving and Inquiry Tasks



Ina V.S. Mullis
Michael O. Martin
Bethany Fishbein
Pierre Foy
Sebastian Moncaleano

© IEA, 2021

IEA
TIMSS & PIRLS
BOSTON COLLEGE

Copyright © 2021
International Association for the Evaluation of Educational Achievement (IEA)

Findings from the TIMSS 2019 Problem Solving and Inquiry Tasks
Ina V.S. Mullis, Michael O. Martin, Bethany Fishbein, Pierre Foy, and Sebastian Moncaleano

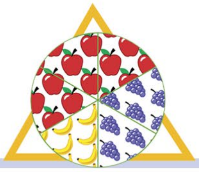
Publishers: TIMSS & PIRLS International Study Center,
Lynch School of Education and Human Development, Boston College and
International Association for the Evaluation of Educational Achievement (IEA)

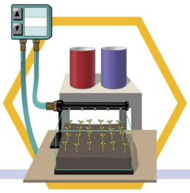
Library of Congress Catalog Card Number: 2021918336
ISBN-978-1-889938-58-5

For more information about TIMSS contact:
TIMSS & PIRLS International Study Center
Lynch School of Education and Human Development
Boston College
Chestnut Hill, MA 02467
United States

tel: +1-617-552-1600
e-mail: timssandpirls.bc.edu
timssandpirls.bc.edu

Boston College is an equal opportunity, affirmative action employer.





FOREWORD

Dirk Hastedt

Executive Director, IEA

TIMSS (the Trends in International Mathematics and Science Study) is the largest and most comprehensive large-scale assessment of mathematics and science for primary and secondary education. It is truly a global enterprise studying the primary and secondary education with more than 70 education systems participating worldwide. TIMSS was first conducted in 1995, and has continued every four years since that date, and as such, has the longest trends of mathematics and science achievement. The richness of TIMSS is not only the availability of achievement measures but also the rich background information collected from the assessed students, their mathematics and science teachers, school principals, parents of the grade four students, as well as system level data. This offers a holistic perspective of the education in the participating countries.

With the 2019 cycle of TIMSS, an additional source of information has been recorded, as the study starts to move to a computer-based assessment format. Out of the education systems participating in TIMSS 2019, 36 of the fourth grade participants and 27 of the eighth grade participants transferred to a computer-based format of the test. This is not only beneficial in terms of matching the mode of the assessment to the reality of teaching and learning in the 21st century, but also allows additional information on how students maneuvered through the test to be captured. New innovative item formats can be administered on a digital device which are engaging and capture aspects of learning that otherwise are very difficult to record. This makes the assessment more enjoyable for participating students, and at the same time creates richer data to be analyzed by researchers.

TIMSS is a flagship study of the International Association for the Evaluation of Educational Achievement (IEA), an independent, international cooperative of national educational research institutions and governmental research agencies dedicated to improving education. IEA's mission is to enhance knowledge about education systems worldwide and to provide high-quality data that will support education reform and lead to better teaching and learning in schools. The TIMSS & PIRLS International Study Center at Boston College has led TIMSS since the first cycle conducted in 1995 and has studied every four years the achievement of fourth and eighth grade students in countries all around the world—always developing the study further while still maintaining reliable and accurate trend measures. TIMSS 2019, the seventh TIMSS cycle, mastered a new challenge by transitioning to computer-based assessment while still maintaining the 24-year-long trend lines.

IEA is extremely proud to work together with this fabulous group of dedicated and recognized researchers.

While conducting TIMSS 2019, the TIMSS & PIRLS International Study Center worked closely with the staff of IEA in their offices in Amsterdam and in Hamburg, as well as Statistics Canada and the Educational Testing Service. Essential for TIMSS—and all other IEA studies—is the close cooperation with the experts from the participating countries who are not only responsible for conducting the study in their respective countries, but also contributing substantially to the study framework, instruments, procedures, and reports. Without these dedicated researchers from around the world, TIMSS surely wouldn't achieve the quality and recognition that it currently has. Additional input was also gathered from specialist committees like the Science and Mathematics Item Review Committee (SMIRC) and the Questionnaire Item Review Committee (QIRC).

This publication is based on the newly developed Problem Solving and Inquiry (PSI) blocks of TIMSS 2019. The report demonstrates the richness and innovativeness of the new item formats, while also highlighting the challenges, and learnings from their administration. The development of the PSI blocks was very labor-intensive and costly, but this report shows what can be learned when moving away from traditional item formats. Consequently, I strongly believe that this report will not only be useful for the further development of TIMSS but will also help other computer-based studies to learn how to improve their assessment.

The work presented in this publication represents the efforts of many individuals and groups. I would like to congratulate the authors of this report, Ina V.S. Mullis, Michael O. Martin, Bethany Fishbein, Pierre Foy, and Sebastian Moncaleano from the TIMSS & PIRLS International Study Center for putting together this informative and valuable publication. I also would like to thank all researchers from the TIMSS & PIRLS International Study Center—especially the Executive Directors Ina V.S. Mullis, Michael O. Martin, and Matthias von Davier, the IEA, Statistics Canada, Educational Testing Service and all the national centers for their vital work on TIMSS—which made this ambitious publication possible. Likewise, without the financial support from the participating countries, the National Center for Education Statistics of the U.S. Department of Education, and the European Commission, the study would not have been possible. Lastly, my deep gratitude goes to the 580,000 students, 52,000 teachers, 19,000 principals, and 310,000 parents who participated in TIMSS 2019—without them there would be no data—and IEA sincerely appreciates and values their willingness to be part of this research project.

Contents

Chapter 3

Building

Pepper Plants

Appendix C

ACKNOWLEDGEMENTS

Countless people made significant contributions to the four-year pioneering effort required to develop the TIMSS 2019 Problem Solving and Inquiry Tasks. Mathematics and science assessment specialists contributed the creative ideas underlying the tasks, in particular, Ray Philpot (ACER), Mary Lindquist (NCTM), Kathleen Scalise (ETS and University of Oregon), and Emily Jones (NFER). The individuals responsible for eTIMSS in the participating countries conducted special pilots, an ambitious field test, and participated in numerous reviews to inform this new endeavor. The TIMSS mathematics and science coordinators, Kerry Cotter and Victoria Centurino, spent many hours on the iterative process of producing and revising the tasks, receiving invaluable support from the international experts comprising the TIMSS 2019 Science and Mathematics Item Review Committee (SMIRC), the talented graphic designers at the TIMSS & PIRLS International Study Center, and dedicated computer specialists at IEA Hamburg.

Introduction

The TIMSS 2019 Problem Solving and Inquiry (PSI) tasks were developed to gain insights into how using digitally-based interactive assessment items to capture students' responses could be incorporated into TIMSS. The goal was ***not*** *to define new problem solving and inquiry constructs, but to collect information that would help enhance and extend the breadth of the TIMSS assessment to provide more comprehensive coverage of problem solving and inquiry as* ***already*** *described in the assessment frameworks.*

TIMSS 2019 Transition to Digital Assessment

In 2019, TIMSS transitioned to digital mathematics and science assessments at both fourth and eighth grades. Half the nearly 70 countries participating in TIMSS 2019 administered the new eTIMSS digitally based assessment, contributing to its development through pilot studies and field testing, while the other half continued with paperTIMSS. By carefully managing how the eTIMSS computerization of items was introduced into TIMSS 2019, making some obvious improvements (e.g., clicking a response to multiple-choice item rather than filling in a circle), but still taking great care to mirror paperTIMSS, most of the items in eTIMSS and paperTIMSS had similar psychometric properties. As documented in *Methods and Procedures: TIMSS 2019 Technical Report*,[1] a complicated step-by-step scaling process enabled linking the eTIMSS and paperTIMSS to the TIMSS mathematics and science achievement scales. The publication of the *TIMSS 2019 International Results in Mathematics and Science*[2] reporting the results for all participating countries on the TIMSS 2019 mathematics and science achievement scales signaled that the transition was complete.

As an important feature of the transition, eTIMSS created the opportunity to develop innovative assessment measures that would enhance coverage of problem solving and inquiry processes. It was evident that a computer-based TIMSS had the potential for improving the quality of the TIMSS measures of higher-order skills (e.g., more depth of concepts, dynamic features, and process data), while at the same time making the data collection of complex tasks feasible. Assessing the TIMSS 2019 frameworks with engaging, computerized assessment tasks benefitting from the most current research became an explicit TIMSS 2019 development goal. Beginning in 2017, the TIMSS & PIRLS International Study Center began developing the "TIMSS 2019 Problem Solving and Inquiry" tasks. Eventually, eight tasks were developed—two for mathematics and two for science each at fourth grade and eighth grade. The eight tasks were assembled into two special eBooklets per grade that were assessed together with eTIMSS in the eTIMSS countries according to a rotated design (see Appendix A). Thus, all the eTIMSS countries (but no paperTIMSS countries) participated in assessing the TIMSS 2019 Problem Solving and Inquiry tasks, including 30 countries and 6 benchmarking systems with about 22,000 students at the fourth grade, and 22 countries and 5 benchmarking systems with about 20,000 students at the eighth grade.

This report presents four of the Problem Solving and Inquiry tasks together with the achievement results across the countries, focusing on the strengths and weaknesses of the tasks themselves.

- ***School Party*—fourth grade mathematics**: Students plan a party for their school (ticket sales, decorations, food, and drinks).
- ***Farm Investigation*—fourth grade science**: A boy investigates which farm animal ate the plants in his garden.
- ***Building*—eighth grade mathematics**: Students construct a storage shed with a rain barrel.
- ***Pepper Plants*—eighth grade science**: Students conduct an experiment to determine the most effective fertilizer.

eTIMSS 2019 also made it possible to collect valuable process data about the ways students proceed through the assessment sessions. This included extensive process data on event timing, navigation from screen to screen, scrolling, and the use of calculators and rulers. These data make it possible to recreate the student's progress through the tasks, and were particularly useful in analyzing non-response data; distinguishing between students who ran out of time and those who stopped responding before time was up. Before erroneously assuming students needed additional assessment time for the PSI tasks, it was important to learn that "running out of time" was less common than "stopping" with plenty of time remaining (see Appendix B). Understanding their reasons for stopping requires further research, but probably some were tired or frustrated. Further highlighting its research potential, the TIMSS 2019 process data also was used for analysis of incorrect responses and learning more about how students dealt with the interactive features to help explain why sometimes performance was lower than expected. Upon discovering considerable non-response to some of the PSI tasks, especially compared to nearly negligible non-response for the "regular" eTIMSS item, the timing data was used to investigate the low completion rates.

Finally, as a byproduct of the PSI tasks, one item in *Building* asked students to show how they would cut the walls out of a board. These responses were used to study the feasibility of TIMSS using automated scoring in the future (see Appendix C). Looking back in time to the TIMSS 2019 transition to digital assessment, the decision to move forward and take advantage of technology and new psychometric research will be recognized as starting a sea change in TIMSS assessment methods and procedures.

Brief History of TIMSS and Problem Solving and Inquiry Tasks

Innovative assessments to assess higher-order skills have been part of TIMSS since its inception. The inaugural TIMSS 1995 included what was at the time considered to be a "state-of-the-art" performance assessment that was given to fourth grade students in 10 countries and eighth grade students in 21 countries. As explained in the TIMSS 1995 report of the results, the performance assessment was based on integrated, practical tasks involving instruments and equipment as a means of assessing students' content and procedural knowledge, as well as their ability to use that knowledge in reasoning and problem solving (see *TIMSS 1995 Performance Assessment*[3]). Performance assessment was considered particularly useful for assessing science as a process of inquiry (beyond just a body of knowledge). Of the 12 tasks given to the fourth and eighth grade students, 11 were similar across grades and one was unique. There were five mathematics tasks—Dice, Calculator, Folding and Cutting, Around the Bend, and Packaging; and five science tasks—Pulse, Magnets, Batteries, Rubber Band, and Containers (fourth grade) or Solutions (eighth grade). Considerable effort was expended in assessing a framework of "performance expectations" that included problem solving, designing an investigation, analyzing and interpreting findings, as well as formulating conclusions.

The performance assessment was administered in a "circus-ring" format where students visited three of five stations located around a room, each consisting of the assembled equipment for one or two tasks. The equipment for the tasks weighed about 100 lbs and needed to be set up in a large room. Thus, it was only feasible to give this very labor and resource intensive assessment to subsamples of students that had participated in the main assessment.

When TIMSS 2003 established regularly administered assessments at the fourth and eighth grades every four years to monitor trends, the U.S. National Science Foundation (NSF) awarded Boston College a grant to support framework and assessment development. The idea was to develop extended problem solving and inquiry tasks, but using only paper-and-pencil instruments. Progress was made on developing content assessment goals tailored specifically to fourth or eighth grade, but the mathematicians, scientists, and measurement community struggled to make the paper-and-pencil tasks accessible to the students as well as engaging. The performance assessment was different and "fun," for example, in a task about the effects of exercise on the body, students got to jump up and down to get their heart rates up. As a disadvantage, students in the TIMSS 2003 participating countries faced an unfamiliar idea—a test that gave you a long time to work through a series of items on a topic (e.g., an ocean food chain or why different colors of light can change the color of your shirt). In general, the early paper-and-pencil PSI tasks of 2003 were not very motivating, so these longer tasks were eventually phased out of upcoming assessments.

Nevertheless, it was widely agreed that the problem solving and inquiry skills were fundamental to the TIMSS assessment frameworks. For TIMSS 2007, the U.S. National Center for Education Statistics (NCES) organized an initiative for countries to contribute funding for TIMSS to develop cognitive as well as content assessment goals. This resulted in three cognitive domains—knowing, applying, and reasoning—becoming a permanent dimension of the mathematics and science assessments at both fourth and eighth grades. Once again for TIMSS 2015, the TIMSS & PIRLS International Study Center at Boston College worked with the National Center for Education Statistics to obtain additional funding from NSF for innovative item development, especially since TIMSS 2015 also included assessing trends in TIMSS Advanced. However, this effort to secure funding was unsuccessful, so the reasoning skills associated with problem solving and inquiry remained in the assessment frameworks with little attention further paid to developing longer assessment tasks. The several problem-solving and inquiry assessment goals shown below have been excerpted from the *TIMSS 2019 Assessment Frameworks*.[4]

Mathematics Frameworks

Reasoning mathematically involves logical, systematic thinking. It includes intuitive and inductive reasoning based on patterns and regularities that can be used to arrive at solutions to problems set in…real life settings.

Determine efficient/appropriate operations, strategies, and tools for solving problems for which there are commonly used methods of solution.

Implement strategies and operations to solve problems involving familiar mathematical concepts and procedures.

Link different elements of knowledge, related representations, and procedures to solve problems.

Science Frameworks

Scientists engage in scientific inquiry by following key science practices that enable them to investigate the natural world and answer questions about it. Students of science must become proficient at these practices…

Use a diagram or other model to demonstrate knowledge of science concepts, to illustrate a process, cycle, relationship, or system, or to find solutions to science problems.

Provide or identify an explanation for an observation or a natural phenomenon using a science concept or principle.

Plan investigations or procedures appropriate for answering scientific questions or testing hypotheses; and describe or recognize the characteristics of well-designed investigations in terms of variables to be measured and controlled and cause-and-effect relationships.

The TIMSS 2019 Problem Solving and Inquiry (PSI) Tasks

Re-imagined for TIMSS 2019, PSI tasks are visually attractive, interactive scenarios that present students with adaptive and responsive ways to follow a series of steps (assessment items) toward a solution or goal. The students' responses are provided via a mixture of selection and constructed response items as well as through various innovative formats to capture students' responses (e.g., number pad, drag and drop, graphing tools, and free drawings).

There are many different ways of instantiating a PSI task. For example, a PSI task can be:

- An interactive science experiment, where students set up and run the experiment, adjusting settings and observing the results (see *Pepper Plants*—Science Eighth Grade).
- A mathematics problem, where students work from a visualization to a finished product involving multiple steps and evaluation of interim results (see *Building*—Mathematics Eighth Grade).
- A mathematical or scientific model that can be manipulated by the students (e.g., predator-prey relationships, solutions, or forces and motion).
- A systematic investigation of the attributes of an object, place, or living organism, implementing a process, or considering cause and effect relationships embedded in a scenario that is compelling and targets topics in the framework (see *School Party* and *Farm Investigation*, Mathematics and Science, respectively, Fourth Grade).

Generally,

- Each PSI task should be situated in a real world, problem, investigation, or activity that provides an underlying narrative or theme for the items. The problem or situation must be sufficiently wide to encompass a number of content and cognitive areas in the Mathematics or Science Frameworks. As much as possible, PSI tasks should attempt to include items addressing various content topics and a range of cognitive demands.
- The narrative should provide a logical or chronological progression from the first item to the ending.
- Because PSI tasks with a single narrative from start to finish can be hard to achieve, PSI tasks also can be written that do not have much narrative, provided there is a common theme to link the items together. The thematic type of PSI task gives students an opportunity to interact with various aspects of a scenario without the order of the interactions having an impact. The items can be independent, while still being coherent and engaging.

In any PSI task, it is important that the items are independent of each other. Whether or not a student gets one item correct should not affect whether the student gets another item correct. That is, in general an answer to an item should not give students a clue so that they could go back and change the answer to a previous item. Or, an item should not be based on a correct answer to the previous item, because not all students will have provided the correct answer. The various incorrect answers can impact the difficulty of the second item or even make it impossible to answer. On the other hand, if designed properly, process data can be used to research "looking back" behaviors as part of students' test-taking strategies.

Developing the Problem Solving and Inquiry Tasks for eTIMSS 2019

TIMSS 2019 PSI task development at fourth and eighth grades adhered to standard TIMSS procedures for ensuring valid measures of the mathematics and science achievement described in the *TIMSS 2019 Assessment Frameworks*.[5] However, developing new and engaging problem contexts with cohesive sets of achievement items necessitated many more rounds of expert review than usual, so staff at the TIMSS & PIRLS International Study Center began collaborating with members of the TIMSS 2019 Science and Mathematics Item Review Committee (SMIRC) in August 2015 to develop the PSI tasks. This was nearly two years before item writing began for the rest of the TIMSS 2019 field test items (April 2017), and involved five additional in-person meetings at Boston College and numerous online reviews.

Cognitive laboratories involving 34 students in the United States (August 2015) provided critical information about the usability of the eTIMSS interface and various innovative item types. SMIRC as a whole focused its first in-depth review of the PSI tasks on the alignment between the tasks and the frameworks, the extent to which the technology in the tasks supported the intended response processes, and the cross-cultural appropriateness of the problem scenarios. Small pilot tests in several eTIMSS countries provided key information at different points in the development process.

The eTIMSS prePilot including a total of 12 PSI tasks was conducted in September 2016 in three English-speaking countries with experience in conducting digital assessments: Australia, Canada, and Singapore. Each country included students with a range of mathematics and science ability in the prePilot, yielding approximately 100 responses per item at both the fourth and eighth grades. The prePilot provided further information about the usability of newly developed item types and students' success in using the eTIMSS interface, as well as estimates of the amount of time it took students to complete each task and the task's approximate difficulty.

National Research Coordinators (NRCs) reviewed the PSI tasks at their 3rd TIMSS 2019 NRC meeting which was held prior to conducting the field test (March 2017) and then reviewed them again after the field test (August 2018) to select the tasks to be included in the eTIMSS 2019 assessment. The NRCs selected eight PSI tasks (four at fourth grade with 50 items and four at eighth grade with 55 items) for the main data collection. The eight tasks covered a range of mathematics and science content domain topics, and consistent with the goal of the PSI tasks to assess higher-order skills, the majority of the items in the PSIs involved applying and reasoning.

Appendix A provides an overview of the parallel assessment designs for paperTIMSS 2019 and for eTIMSS 2019. The eTIMSS design also specifies the rotated arrangement of the eight PSI tasks—two for mathematics and two for science at each grade. Both fourth and eighth grades included two separate booklets of PSI items.

Including the PSI items in the TIMSS 2019 Mathematics and Science Achievement Scales at Fourth and Eighth Grades

Exhibits 1 through 4 compare TIMSS 2019 achievement estimated with and without the PSI data for the eTIMSS countries (one exhibit each for mathematics at fourth grade, science at fourth grade, mathematics at eighth grade, and science at eighth grade, respectively). The first column in each exhibit is a reproduction of the average achievement results published in *TIMSS 2019 International Results in Mathematics and Science*[6] for countries that administered the digital version of TIMSS (eTIMSS). The second column presents the average achievement results for eTIMSS including the TIMSS 2019 PSIs (for details of the scaling procedures, see Chapter 17 in *Methods and Procedures: TIMSS 2019 Technical Report*[7]). For each grade, there essentially was no difference (0 scale score points on average) between eTIMSS average achievement excluding the PSI students compared to average achievement including the PSI students for either mathematics or science.

Exhibit 1

Average Mathematics Achievement for eTIMSS Compared to eTIMSS with PSI—Fourth Grade

Country	eTIMSS Mathematics Average Scale Score (not including PSI)	eTIMSS with PSI Mathematics Average Scale Score	Difference
[3] Singapore	625 (3.9)	623 (3.8)	-3 (0.3)
† Hong Kong SAR	602 (3.3)	601 (3.3)	-1 (0.5)
Korea, Rep. of	600 (2.2)	599 (2.2)	-1 (0.5)
Chinese Taipei	599 (1.9)	598 (1.8)	-1 (0.4)
[2] Russian Federation	567 (3.3)	567 (3.2)	0 (0.4)
[2] England	556 (3.0)	557 (2.8)	1 (0.6)
† Norway (5)	543 (2.2)	544 (2.2)	1 (0.5)
[2] Lithuania	542 (2.8)	542 (2.8)	0 (0.4)
Austria	539 (2.0)	539 (2.0)	0 (0.3)
≡ Netherlands	538 (2.2)	539 (2.1)	1 (0.4)
[2] † United States	535 (2.5)	534 (2.5)	-1 (0.3)
Czech Republic	533 (2.5)	533 (2.4)	0 (0.5)
Finland	532 (2.3)	532 (2.2)	0 (0.5)
[2] Portugal	525 (2.6)	524 (2.6)	-1 (0.4)
† Denmark	525 (1.9)	526 (1.9)	2 (0.4)
Hungary	523 (2.6)	524 (2.6)	1 (0.4)
[2] Turkey (5)	523 (4.4)	521 (4.5)	-1 (0.4)
Sweden	521 (2.8)	522 (2.8)	1 (0.5)
Germany	521 (2.3)	521 (2.2)	0 (0.5)
Italy	515 (2.4)	514 (2.3)	0 (0.4)
[1 2] Canada	512 (1.9)	512 (1.8)	0 (0.2)
[2] Slovak Republic	510 (3.5)	511 (3.3)	2 (0.4)
Croatia	509 (2.2)	510 (2.1)	1 (0.4)
Malta	509 (1.4)	509 (1.4)	0 (0.6)
Spain	502 (2.1)	503 (2.1)	1 (0.6)
France	485 (3.0)	485 (3.0)	0 (0.5)
[1] Georgia	482 (3.7)	482 (3.6)	0 (0.6)
United Arab Emirates	481 (1.7)	480 (1.7)	-1 (0.2)
Qatar	449 (3.4)	449 (3.4)	0 (0.5)
Chile	441 (2.7)	442 (2.7)	1 (0.4)
International Average	**528 (0.5)**	**528 (0.5)**	**0 (0.1)**
Benchmarking Participants			
Moscow City, Russian Fed.	593 (2.2)	592 (2.1)	0 (0.5)
[2] Dubai, UAE	544 (1.6)	543 (1.6)	-1 (0.4)
Quebec, Canada	532 (2.3)	531 (2.2)	-1 (0.5)
Madrid, Spain	518 (2.2)	519 (2.0)	1 (0.5)
[2] Ontario, Canada	512 (3.3)	512 (3.2)	0 (0.4)
Abu Dhabi, UAE	441 (2.2)	440 (2.3)	-1 (0.3)

See Chapter 9 in *Methods and Procedures: TIMSS 2019 Technical Report* for population coverage notes 1, 2, and 3 and for sampling participation notes †, ‡, and ≡.

( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Exhibit 2

Average Science Achievement for eTIMSS Compared to eTIMSS with PSI—Fourth Grade

Country	eTIMSS Science Average Scale Score (not including PSI)	eTIMSS with PSI Science Average Scale Score	Difference
[3] Singapore	595 (3.4)	593 (3.4)	-1 (0.3)
Korea, Rep. of	588 (2.1)	588 (2.1)	0 (0.5)
[2] Russian Federation	567 (3.0)	567 (2.9)	0 (0.4)
Chinese Taipei	558 (1.8)	557 (1.8)	-1 (0.5)
Finland	555 (2.6)	554 (2.5)	-1 (0.4)
[†] Norway (5)	539 (2.2)	540 (2.3)	1 (0.5)
[2 †] United States	539 (2.7)	538 (2.7)	-1 (0.4)
[2] Lithuania	538 (2.5)	538 (2.5)	0 (0.4)
Sweden	537 (3.3)	538 (3.3)	0 (0.5)
[2] England	537 (2.7)	539 (2.6)	2 (0.5)
Czech Republic	534 (2.6)	534 (2.5)	0 (0.5)
[†] Hong Kong SAR	531 (3.3)	531 (3.2)	-1 (0.5)
Hungary	529 (2.7)	529 (2.5)	0 (0.5)
[2] Turkey (5)	526 (4.2)	526 (4.2)	-1 (0.4)
Croatia	524 (2.2)	524 (2.1)	0 (0.5)
[1 2] Canada	523 (1.9)	523 (1.9)	0 (0.3)
[†] Denmark	522 (2.4)	524 (2.3)	2 (0.5)
Austria	522 (2.6)	522 (2.5)	0 (0.4)
[2] Slovak Republic	521 (3.7)	522 (3.6)	1 (0.6)
[≡] Netherlands	518 (2.9)	520 (2.8)	1 (0.4)
Germany	518 (2.2)	519 (2.1)	0 (0.5)
Spain	511 (2.0)	512 (1.9)	1 (0.5)
Italy	510 (3.0)	509 (2.9)	0 (0.4)
[2] Portugal	504 (2.6)	504 (2.6)	0 (0.4)
Malta	496 (1.3)	496 (1.1)	1 (0.7)
France	488 (3.0)	488 (2.9)	1 (0.4)
United Arab Emirates	473 (2.1)	470 (2.1)	-3 (0.2)
Chile	469 (2.6)	470 (2.6)	1 (0.4)
[1] Georgia	454 (3.9)	455 (3.9)	0 (0.6)
Qatar	449 (3.9)	448 (3.8)	-1 (0.5)
International Average	**523 (0.5)**	**523 (0.5)**	**0 (0.1)**
Benchmarking Participants			
Moscow City, Russian Fed.	595 (2.2)	594 (2.1)	-1 (0.4)
[2] Dubai, UAE	545 (1.7)	542 (1.7)	-3 (0.4)
[2] Ontario, Canada	524 (3.2)	524 (3.1)	0 (0.5)
Madrid, Spain	523 (2.0)	523 (1.9)	0 (0.5)
Quebec, Canada	522 (2.5)	522 (2.4)	0 (0.5)
Abu Dhabi, UAE	418 (2.8)	416 (2.8)	-2 (0.4)

See Chapter 9 in *Methods and Procedures: TIMSS 2019 Technical Report* for population coverage notes 1, 2, and 3 and for sampling participation notes †, ‡, and ≡.
( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Exhibit 3

Average Mathematics Achievement for eTIMSS Compared to eTIMSS with PSI—Eighth Grade

Country	eTIMSS Mathematics Average Scale Score (not including PSI)	eTIMSS with PSI Mathematics Average Scale Score	Difference
[2] Singapore	616 (4.0)	616 (3.9)	0 (0.3)
Chinese Taipei	612 (2.7)	610 (2.7)	-3 (0.5)
Korea, Rep. of	607 (2.8)	604 (2.7)	-3 (0.6)
† Hong Kong SAR	578 (4.1)	579 (4.1)	0 (0.4)
[2] Russian Federation	543 (4.5)	544 (4.5)	0 (0.4)
Lithuania	520 (2.9)	521 (2.9)	0 (0.5)
[3] Israel	519 (4.3)	518 (4.3)	-1 (0.4)
Hungary	517 (2.9)	517 (2.9)	1 (0.5)
† United States	515 (4.8)	516 (4.7)	1 (0.4)
England	515 (5.3)	515 (5.1)	0 (0.5)
Finland	509 (2.6)	509 (2.6)	0 (0.4)
† Norway (9)	503 (2.4)	504 (2.4)	1 (0.5)
[2] Sweden	503 (2.5)	504 (2.6)	2 (0.5)
Portugal	500 (3.2)	501 (3.1)	0 (0.5)
Italy	497 (2.7)	497 (2.8)	0 (0.5)
Turkey	496 (4.3)	495 (4.1)	-1 (0.6)
France	483 (2.5)	484 (2.4)	1 (0.5)
United Arab Emirates	473 (1.9)	474 (1.9)	0 (0.2)
[1] Georgia	461 (4.3)	460 (4.2)	-1 (0.6)
Malaysia	461 (3.2)	462 (3.2)	1 (0.4)
Ψ Qatar	443 (4.0)	443 (4.0)	0 (0.6)
Ψ Chile	441 (2.8)	441 (2.7)	0 (0.7)
International Average	**514 (0.7)**	**514 (0.7)**	**0 (0.1)**
Benchmarking Participants			
Moscow City, Russian Fed.	575 (4.2)	575 (4.1)	0 (0.5)
‡ Quebec, Canada	543 (3.7)	544 (3.6)	1 (0.4)
[2] Dubai, UAE	537 (2.0)	537 (2.0)	1 (0.4)
Ontario, Canada	530 (4.3)	531 (4.3)	2 (0.5)
Ψ Abu Dhabi, UAE	436 (2.9)	436 (3.0)	0 (0.3)

Ψ Reservations about reliability because the percentage of students with achievement too low for estimation exceeds 15% but does not exceed 25%.

See Chapter 9 in *Methods and Procedures: TIMSS 2019 Technical Report* for population coverage notes 1, 2, and 3 and for sampling participation notes †, ‡, and ≡.

( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Exhibit 4

Average Science Achievement for eTIMSS Compared to eTIMSS with PSI—Eighth Grade

Country	eTIMSS Science Average Scale Score (not including PSI)	eTIMSS with PSI Science Average Scale Score	Difference
[2] Singapore	608 (3.9)	608 (3.9)	1 (0.3)
Chinese Taipei	574 (1.9)	573 (1.9)	-1 (0.4)
Korea, Rep. of	561 (2.1)	560 (2.0)	0 (0.6)
[2] Russian Federation	543 (4.2)	542 (4.1)	-1 (0.4)
Finland	543 (3.1)	543 (3.0)	0 (0.4)
Lithuania	534 (3.0)	534 (3.0)	0 (0.4)
Hungary	530 (2.6)	530 (2.6)	0 (0.4)
† United States	522 (4.7)	525 (4.5)	3 (0.5)
[2] Sweden	521 (3.2)	523 (3.3)	1 (0.6)
Portugal	519 (2.9)	518 (2.8)	-1 (0.4)
England	517 (4.8)	518 (4.7)	1 (0.5)
Turkey	515 (3.7)	514 (3.6)	-1 (0.6)
[3] Israel	513 (4.2)	513 (4.2)	0 (0.5)
† Hong Kong SAR	504 (5.2)	506 (5.2)	2 (0.5)
Italy	500 (2.6)	499 (2.6)	-2 (0.5)
† Norway (9)	495 (3.1)	497 (3.1)	1 (0.6)
France	489 (2.7)	491 (2.6)	2 (0.5)
Qatar	475 (4.4)	473 (4.3)	-2 (0.5)
United Arab Emirates	473 (2.2)	471 (2.2)	-2 (0.3)
Chile	462 (2.9)	463 (2.8)	1 (0.6)
Malaysia	460 (3.5)	461 (3.5)	1 (0.3)
[1] Georgia	447 (3.9)	446 (3.7)	0 (0.7)
International Average	**514 (0.7)**	**514 (0.7)**	**0 (0.1)**
Benchmarking Participants			
Moscow City, Russian Fed.	567 (2.9)	567 (2.9)	0 (0.5)
[2] Dubai, UAE	548 (2.0)	546 (2.0)	-2 (0.4)
‡ Quebec, Canada	537 (3.6)	538 (3.6)	1 (0.4)
Ontario, Canada	522 (3.0)	524 (2.9)	3 (0.5)
Abu Dhabi, UAE	420 (3.6)	419 (3.6)	-1 (0.4)

See Chapter 9 in *Methods and Procedures: TIMSS 2019 Technical Report* for population coverage notes 1, 2, and 3 and for sampling participation notes †, ‡, and ≡.

( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Important Information for Future Development

It should be noted that the concept of an effective PSI task will continue to evolve, because following publication of this report at the end of October, IEA will release the process data for the TIMSS 2019 PSI tasks, enabling a series of further in-depth analyses. Basic criteria that were important in TIMSS 2019 remain, however, additional considerations have emerged:

- The PSI task must address topics in the TIMSS mathematics or science frameworks.
- PSI tasks can be full length at eighth grade (a block of 10 to 15 items) or "mini" about 5 to 8 items. At fourth grade, there only will be mini-PSI tasks in TIMSS 2023.
  - The completion rates presented in Appendix A for the items in the eTIMSS assessment compared to the PSI tasks show that the fourth grade PSI tasks had comparatively low completion rates.
- No PSI task or items should require excessive reading, perseverance, or specialized knowledge.
- Typically, the first screen introduces the topic, and the following screens present the items (no ending screen).
- Each PSI task should include a range of item difficulty. Typically a task should start with easier items and end with more difficult items.
- PSI items should not be dependent on other items (unless it is for planned research purposes).
- PSI items should take advantage of the digital environment, using interactive or adaptive features, but not gratuitously.
- The mode of capturing the students' responses should assist the students in displaying their mathematics or science understanding, not create a distraction.
- PSI tasks and items should be designed to capitalize the potential of process data.
- PSI items must adhere to the *TIMSS 2019 Item Writing Guidelines*.[8]
- A scoring guide needs to accompany each human scored PSI item. (Partial credit may be awarded if warranted. Process data may be used for this purpose.)

Notes

[1] Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 12.1–12.146). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html

[2] Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

[3] Harmon, E., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez, E. J., & Orpwood, G. (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study (TIMSS)*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timss.bc.edu/timss1995i/PAreport.html

[4] Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/

[5] Ibid.

[6] Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

[7] Fishbein, B., & Foy, P. (2021). Scaling the TIMSS 2019 problem solving and inquiry data. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 17.1–17.51). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-17.html

[8] Mullis, I. V. S., Martin, M. O., Cotter, K. E., & Centurino, V. A. S. (2020). *TIMSS 2019 Item Writing Guidelines*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/pdf/T19-item-writing-guidelines.pdf

CHAPTER 1

Mathematics Grade 4

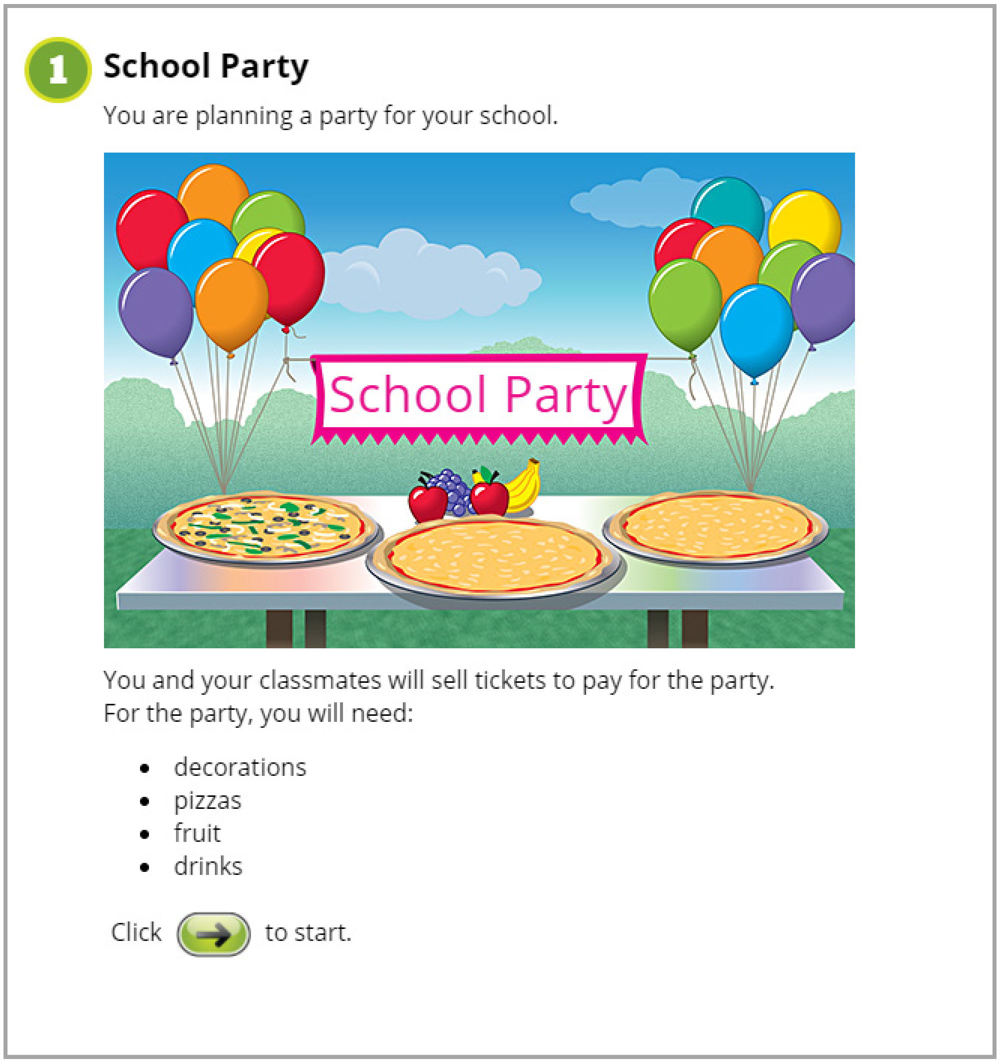
School Party

About the Task

In the *School Party* PSI task, fourth grade students were asked to plan a party for their school. Based on attendance at last year's school party, they were asked to plan the party for 400 people. The party planning involved considering the price of the tickets and what decorations, food, and drinks to purchase for the party. The task is colorful and has a moderate degree of interactivity to keep students engaged. However, compared to the eTIMSS items, it had comparatively high levels of non-response for the last items (see Appendix A for details).

Screen 1 – Introduction

The task was presented in a series of eight screens, with Screen 1 (shown below) introducing the task. Screen 1 does not have any items. Following Screen 1 (introductory), Screens 2 through 8 (each with one or two items) guide the students through 12 mathematics items related to the party planning activities.



1 **School Party**

You are planning a party for your school.

You and your classmates will sell tickets to pay for the party.
For the party, you will need:

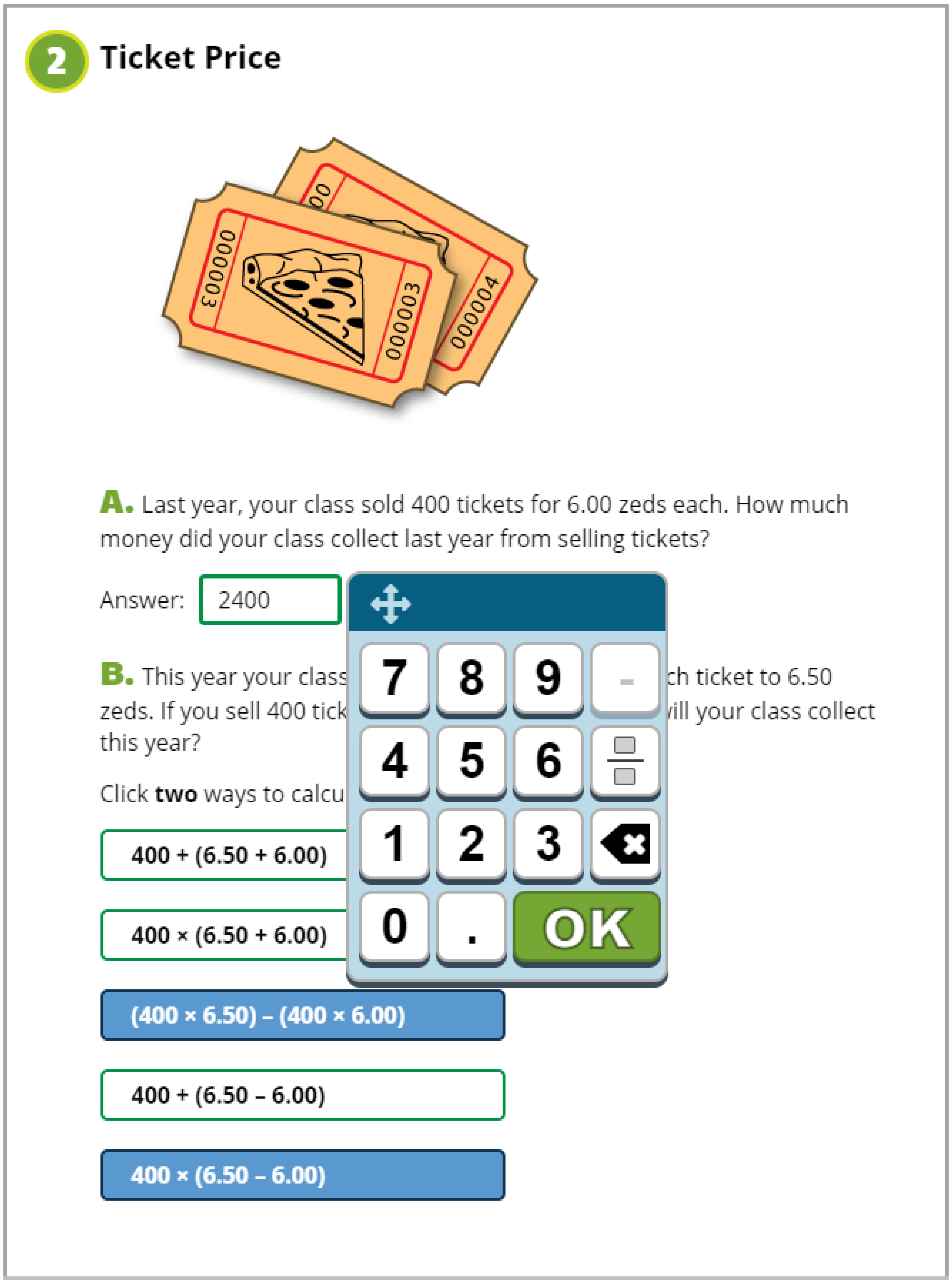
- decorations
- pizzas
- fruit
- drinks

Click ➔ to start.

Screen 2 – Ticket Price

It is good practice to begin sets of assessment items with problems accessible to students so they can gain confidence in their ability to continue. Because the PSIs typically have a series of related problems, it is even more important for students not to “become lost” during the first part of the PSI. Most of the fourth grade students, 91 percent, engaged with 2A of the *School Party* task, which was a relatively straightforward multiplication problem. Students were asked to determine the amount of money the previous year’s school party had raised by selling 400 tickets that cost 6.00 zeds for each ticket, with the correct answer 2400 or equivalent. (A zed is one unit of the fictitious currency used since 1995 in TIMSS items involving money to provide the same level of difficulty across countries.)

For all the items with numerical answers, students at both grades entered their responses into the green boxes using the TIMSS number pad (shown below).



2 Ticket Price

A. Last year, your class sold 400 tickets for 6.00 zeds each. How much money did your class collect last year from selling tickets?

Answer: 2400

B. This year your class ch ticket to 6.50 zeds. If you sell 400 tick vill your class collect this year?

Click **two** ways to calcu

400 + (6.50 + 6.00)

400 × (6.50 + 6.00)

(400 × 6.50) – (400 × 6.00)

400 + (6.50 – 6.00)

400 × (6.50 – 6.00)

	Item 2A	Item 2B
Maximum Score Points:	1	2
Content Domain:	Number	Number
Topic Area:	Whole Numbers	Expressions, Simple Equations, and Relationships
Cognitive Domain:	Applying	Applying

The results for 2A are shown in Exhibit 5, which has the percent of correct responses given by students in each of the eTIMSS countries from highest to lowest. Led by Hong Kong SAR, more than half the students (55 to 79%) in 7 countries provided the correct answer. However, the average across the 30 eTIMSS countries was 42 percent. TIMSS has shown that fourth grade students sometimes find computation with money difficult because of the decimals, and further analysis of the incorrect responses revealed that 9 percent answered 24, 240, 24000, or 240000. Also, 9 percent on average across countries, often Nordic or European countries, omitted this item. However, no other patterns appeared from searching through the remaining incorrect responses. There were a few students entering 4 or 6 or both (e.g., 406) into the number pad, but it is difficult to interpret whether they were trying to add to solve the problem, perhaps trying to use the number pad as a calculator, or just did not understand what was being assessed. Across the participating countries, on average, boys had a higher percent of correct responses than girls.

School Party Screen 2A – Percent Correct Overall and by Gender

Country	Percent Correct (2400 zeds)		
	Overall Country	Girls	Boys
Hong Kong SAR	79 (2.7)	78 (2.8)	80 (4.3)
Korea, Rep. of	73 (2.1)	71 (2.8)	76 (3.1)
Russian Federation	62 (2.3)	63 (3.1)	62 (2.8)
Chinese Taipei	59 (2.6)	54 (3.8)	63 (3.5)
Czech Republic	55 (2.8)	53 (4.1)	57 (3.4)
Croatia	55 (2.7)	52 (4.3)	57 (3.9)
Singapore	55 (2.0)	50 (3.0)	59 (2.6)
Netherlands	48 (2.1)	45 (3.4)	50 (3.2)
England	47 (2.4)	43 (3.6)	50 (3.5)
Austria	45 (2.5)	41 (4.4)	48 (3.3)
Norway (5)	45 (2.6)	43 (3.8)	46 (3.8)
Georgia	43 (2.6)	40 (3.4)	47 (3.4)
Italy	42 (2.5)	36 (3.8)	47 (3.4)
Lithuania	42 (2.8)	39 (4.0)	44 (3.4)
Sweden	39 (2.8)	35 (3.7)	43 (3.5)
Turkey (5)	39 (2.3)	35 (3.1)	42 (3.8)
Germany	39 (2.5)	36 (3.7)	41 (3.1)
Spain	37 (2.1)	34 (4.1)	40 (3.3)
Hungary	36 (2.1)	33 (2.9)	39 (3.3)
Denmark	36 (2.5)	34 (3.4)	38 (3.8)
Portugal	35 (2.3)	34 (2.9)	37 (2.9)
United States	35 (1.7)	33 (2.3)	38 (2.3)
France	35 (2.2)	36 (3.2)	33 (3.2)
Slovak Republic	34 (2.3)	27 (2.6)	41 (3.5)
Finland	30 (2.1)	26 (2.5)	34 (3.2)
United Arab Emirates	27 (0.8)	26 (1.2)	29 (1.0)
Malta	26 (2.1)	22 (2.7)	30 (3.0)
Canada	26 (1.2)	23 (1.8)	28 (1.8)
Qatar	21 (1.6)	22 (2.1)	21 (2.5)
Chile	20 (1.8)	18 (2.8)	21 (2.6)
International Average	**42 (0.4)**	**39 (0.6)**	**45 (0.6)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	68 (2.5)	69 (3.1)	66 (3.4)
Dubai, UAE	43 (1.6)	43 (2.5)	43 (1.9)
Madrid, Spain	39 (2.2)	36 (3.0)	43 (3.3)
Quebec, Canada	33 (2.8)	34 (3.7)	32 (3.7)
Ontario, Canada	23 (2.0)	17 (3.1)	28 (2.8)
Abu Dhabi, UAE	18 (1.2)	15 (1.9)	20 (1.6)

Percentage significantly higher than other gender ▲

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

2B asked students to consider how much more revenue would be provided if the ticket price was raised to 6.50 zeds. However, students did not need to calculate the actual answer. Instead, to assess the "prealgebra" topic in the number content area in the TIMSS 2019 Mathematics Framework, fourth grade students were given five expressions and asked to identify which two showed a way to calculate the answer.

Exhibit 6 shows the percentages of correct responses given by students in each of the eTIMSS countries from highest to lowest. Except in Singapore, less than half the students were able to identify both (full credit) or one (partial credit) of the ways to calculate the answer. Across countries, 30 percent on average received at least partial credit. The Singaporean fourth grade students posted the highest percent of fully correct responses—35 percent. In the remaining countries, 26 percent or less selected both of the two correct expressions. Another 16 percent of the students, on average, were able to identify one of the correct expressions. As might be anticipated, more students (11%) recognized (400 × 6.50) – (400 × 6.00) as correct but not 400 × (6.50 – 6.00), than vice versa, with very few (5%) recognizing the simplification. There was little difference in achievement between girls and boys.

Exhibit 6

School Party Screen 2B – Percent Full Credit Overall and by Gender

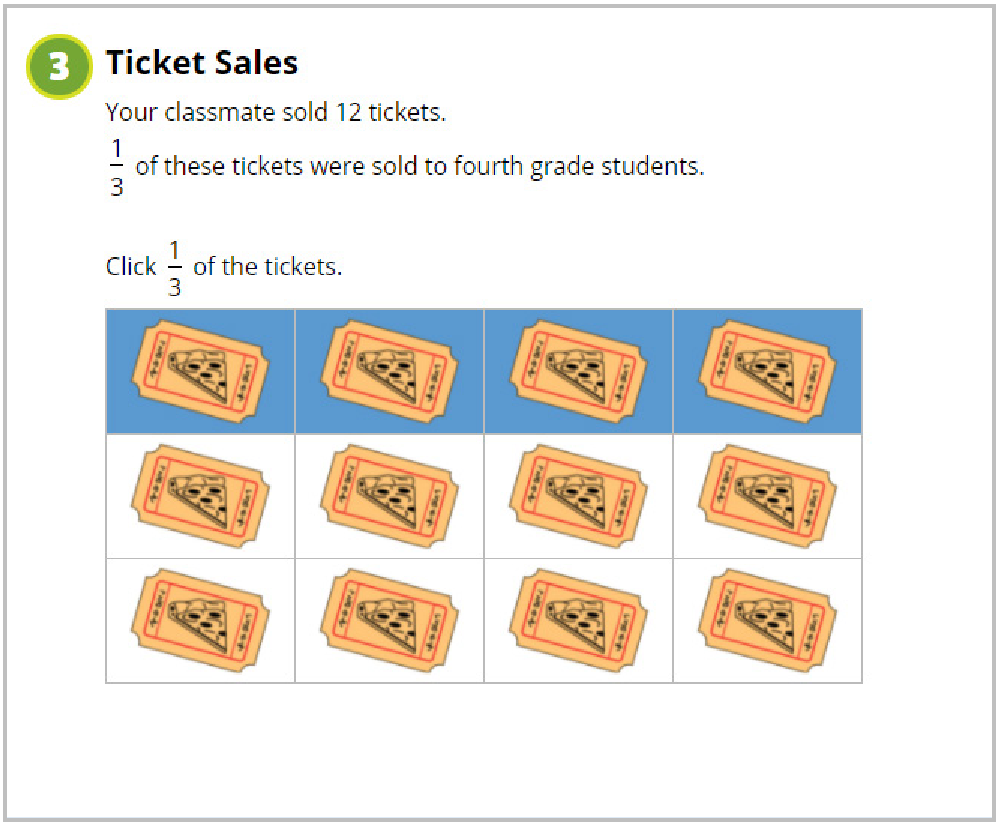
Country	Percent Full Credit (Selects Both Correct Ways)			Percent Partial Credit (Selects Only 1 Correct Way)	
	Overall Country	Girls	Boys	(400 x 6.50) – (400 x 6.00)	400 x (6.50 – 6.00)
Singapore	35 (1.9)	34 (2.7)	36 (2.4)	13 (1.3)	7 (0.9)
Korea, Rep. of	26 (2.1)	22 (3.1)	29 (3.3)	8 (1.2)	1 (0.5)
Hong Kong SAR	25 (2.6)	20 (3.3)	29 (3.8)	15 (1.9)	9 (1.6)
Russian Federation	24 (1.6)	25 (2.5)	23 (1.8)	13 (1.5)	6 (1.1)
Chinese Taipei	22 (2.0)	23 (2.9)	20 (2.5)	10 (1.3)	13 (1.4)
Turkey (5)	19 (1.7)	22 (2.6)	17 (2.5)	10 (1.2)	3 (0.7)
England	17 (1.6)	15 (2.4)	19 (2.7)	10 (1.4)	6 (1.4)
Norway (5)	17 (2.0)	17 (3.4)	16 (3.0)	8 (1.4)	5 (1.2)
Lithuania	15 (1.7)	15 (2.2)	15 (2.6)	13 (1.8)	6 (1.1)
Netherlands	15 (1.9)	14 (3.1)	15 (2.2)	13 (1.9)	4 (1.1)
Sweden	14 (2.0)	13 (2.2)	15 (3.2)	9 (1.5)	5 (1.0)
Finland	14 (1.4)	13 (2.1)	14 (2.2)	9 (1.2)	4 (0.8)
Croatia	14 (2.1)	13 (3.5)	15 (2.3)	4 (1.1)	3 (0.7)
United States	13 (1.1)	12 (1.6)	14 (1.4)	13 (1.1)	7 (0.9)
Italy	13 (1.6)	12 (2.4)	13 (1.9)	14 (1.5)	5 (1.1)
Czech Republic	12 (1.5)	12 (2.1)	13 (2.2)	7 (1.1)	3 (0.9)
Germany	12 (1.5)	14 (2.5)	11 (2.1)	8 (1.2)	3 (0.8)
Austria	12 (1.4)	10 (1.7)	14 (2.1)	11 (1.6)	3 (0.8)
Spain	12 (1.4)	11 (1.3)	13 (2.3)	10 (1.3)	5 (0.8)
Georgia	11 (1.7)	10 (2.7)	11 (2.1)	15 (1.8)	4 (0.8)
Malta	11 (1.4)	10 (2.0)	11 (2.0)	11 (1.3)	4 (0.9)
Portugal	10 (1.5)	10 (2.1)	11 (1.7)	10 (1.6)	6 (1.2)
Denmark	10 (1.6)	14 (2.6)	6 (1.6)	11 (1.7)	2 (0.8)
Canada	10 (1.3)	10 (1.4)	10 (2.0)	13 (1.1)	4 (0.6)
Slovak Republic	9 (1.2)	7 (1.6)	11 (1.8)	11 (1.7)	5 (0.9)
Hungary	9 (1.1)	9 (1.6)	9 (1.8)	12 (1.5)	8 (1.3)
France	9 (1.2)	7 (1.5)	10 (2.1)	8 (1.4)	3 (0.7)
United Arab Emirates	8 (0.6)	7 (0.9)	8 (0.8)	14 (0.6)	6 (0.4)
Qatar	6 (1.2)	6 (1.8)	6 (1.5)	14 (1.4)	4 (0.8)
Chile	3 (0.8)	2 (0.8)	5 (1.3)	8 (1.3)	2 (0.7)
International Average	**14 (0.3)**	**14 (0.4)**	**15 (0.4)**	**11 (0.3)**	**5 (0.2)**
Benchmarking Participants					
Moscow City, Russian Fed.	33 (2.2)	30 (2.9)	37 (3.2)	13 (1.4)	3 (0.8)
Madrid, Spain	16 (2.2)	18 (3.3)	14 (2.2)	8 (1.5)	5 (1.1)
Dubai, UAE	13 (1.0)	10 (1.4)	15 (1.8)	15 (1.3)	7 (0.8)
Quebec, Canada	11 (1.7)	10 (2.2)	11 (2.4)	10 (1.7)	6 (1.0)
Ontario, Canada	11 (2.4)	11 (2.5)	10 (3.6)	16 (2.0)	4 (1.0)
Abu Dhabi, UAE	5 (0.8)	4 (1.0)	6 (0.9)	13 (1.1)	6 (0.8)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 3 – Ticket Sales

Nearly all students tackled the item on Screen 3 assessing fourth grade students' familiarity with basic fractions.



3 **Ticket Sales**

Your classmate sold 12 tickets.

$\frac{1}{3}$ of these tickets were sold to fourth grade students.

Click $\frac{1}{3}$ of the tickets.

Maximum Score Points: 1
Content Domain: Number
Topic Area: Fractions and Decimals
Cognitive Domain: Applying

As shown in Exhibit 7, Singapore and Hong Kong SAR had the highest achievement, with 80–81 percent of their students correctly clicking on 4 of the 12 tickets. On average, nearly half the eTIMSS students answered correctly (48%). One quarter of the students on average answered 3 tickets, possibly indicating a misconception about the fraction $\frac{1}{3}$. On average across countries, there was a gender gap in achievement favoring boys.

Exhibit 7

School Party Screen 3 – Percent Correct Overall and by Gender

Country	Percent Correct (Selects 4 tickets)			Percent Selects Only 3 Tickets
	Overall Country	Girls	Boys	
Hong Kong SAR	81 (2.4)	83 (2.8)	80 (3.1)	9 (1.6)
Singapore	80 (1.6)	79 (2.3)	81 (1.7)	10 (1.0)
Korea, Rep. of	72 (2.3)	69 (3.6)	74 (2.8)	10 (1.4)
Chinese Taipei	65 (2.1)	61 (3.1)	68 (2.7)	13 (1.4)
Denmark	60 (2.5)	58 (3.7)	62 (3.2)	20 (2.1)
Finland	59 (2.0)	58 (2.8)	60 (3.1)	19 (1.6)
Lithuania	59 (2.7)	57 (3.3)	60 (3.6)	28 (2.5)
Norway (5)	59 (3.0)	55 (4.1)	63 (3.9)	19 (2.0)
Netherlands	54 (2.5)	48 (3.7)	61 (3.3)	25 (2.6)
England	52 (2.4)	49 (3.6)	54 (3.0)	29 (2.2)
Canada	51 (1.6)	48 (2.5)	54 (2.2)	25 (1.3)
Malta	50 (2.2)	48 (3.5)	51 (2.8)	25 (1.9)
Sweden	49 (2.8)	49 (4.0)	49 (3.6)	27 (2.0)
Russian Federation	49 (3.0)	45 (3.3)	52 (3.4)	29 (2.4)
Portugal	45 (2.4)	39 (3.4)	51 (3.1)	33 (2.2)
Turkey (5)	45 (2.3)	40 (2.9)	49 (3.4)	25 (1.9)
United States	44 (1.8)	37 (2.5)	50 (2.5)	24 (1.4)
Czech Republic	43 (3.1)	42 (3.8)	44 (4.1)	37 (3.0)
Hungary	43 (2.2)	39 (3.2)	47 (3.0)	32 (2.1)
Slovak Republic	43 (2.5)	39 (3.1)	46 (3.5)	31 (2.5)
United Arab Emirates	39 (1.0)	38 (1.4)	41 (1.5)	26 (0.8)
Italy	39 (2.7)	30 (3.6)	48 (3.3)	26 (2.5)
Austria	37 (2.3)	33 (3.3)	41 (3.1)	30 (1.8)
Croatia	34 (2.8)	35 (4.7)	32 (3.1)	30 (2.3)
Georgia	34 (2.4)	29 (3.6)	37 (3.0)	32 (2.5)
Qatar	33 (2.5)	31 (3.3)	36 (3.5)	29 (2.1)
France	33 (2.6)	28 (3.0)	39 (3.2)	32 (2.0)
Spain	31 (2.2)	25 (2.7)	36 (3.0)	35 (2.0)
Chile	25 (2.1)	26 (2.8)	24 (3.0)	29 (2.2)
Germany	24 (2.2)	22 (3.2)	25 (3.0)	32 (2.3)
International Average	**48 (0.4)**	**45 (0.6)**	**51 (0.6)** ▲	**26 (0.4)**
Benchmarking Participants				
Quebec, Canada	79 (1.9)	81 (2.7)	76 (3.0)	12 (1.6)
Moscow City, Russian Fed.	59 (2.3)	57 (3.3)	61 (3.3)	27 (1.9)
Dubai, UAE	49 (1.8)	45 (2.4)	52 (2.7)	26 (1.8)
Ontario, Canada	42 (2.5)	35 (3.9)	49 (3.3)	29 (2.6)
Madrid, Spain	35 (2.4)	37 (3.3)	33 (3.3)	35 (2.5)
Abu Dhabi, UAE	34 (1.4)	33 (2.2)	36 (1.9)	25 (1.4)

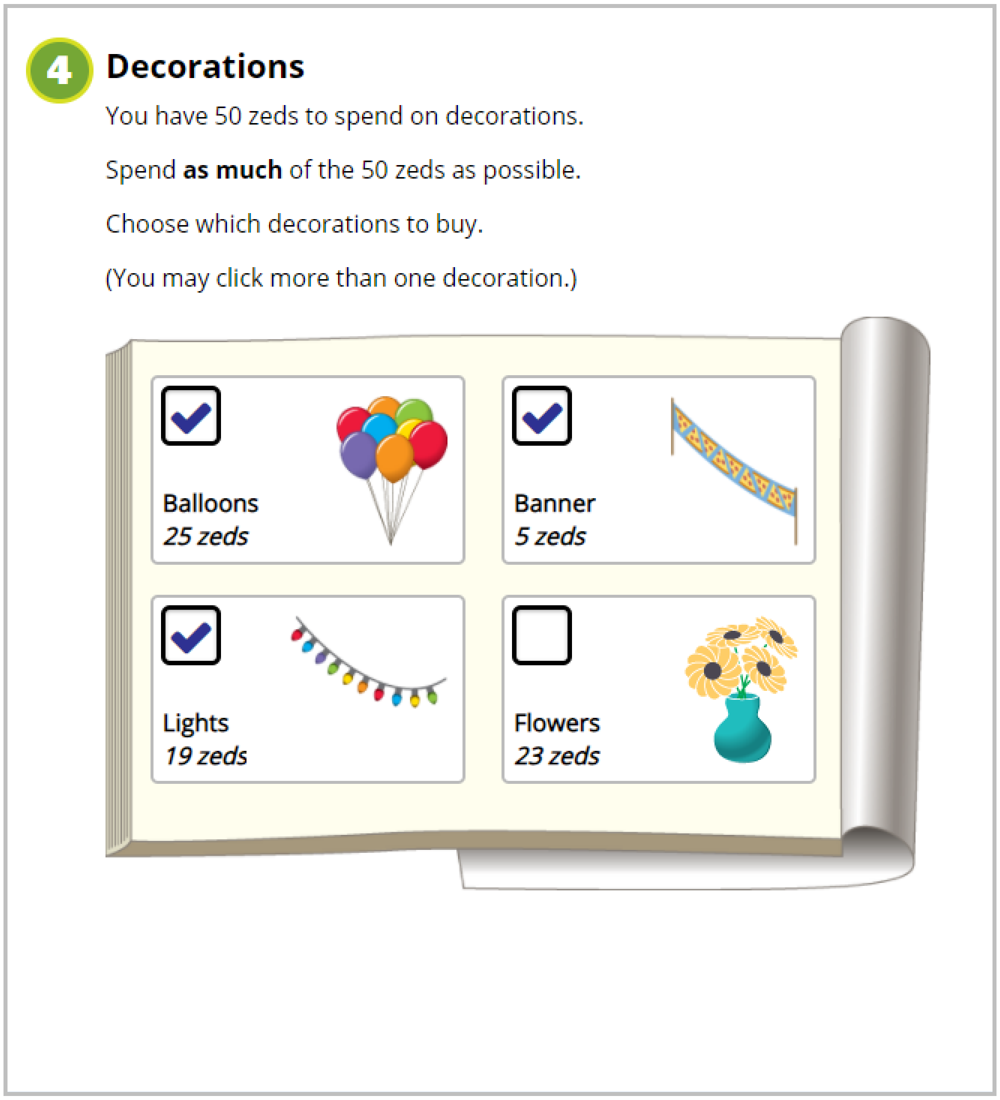
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 4 – Decorations

Nearly all the students (99% on average) responded to the question on Screen 4, which involved reasoning with whole numbers. As shown below, students were given a 50 zed budget for decorations and asked to spend as much of it as possible. The correct answer awarded full credit (2 points) for balloons, lights, and the banner for a total of 49 zeds.



4 **Decorations**

You have 50 zeds to spend on decorations.

Spend **as much** of the 50 zeds as possible.

Choose which decorations to buy.

(You may click more than one decoration.)

☑ Balloons *25 zeds*

☑ Banner *5 zeds*

☑ Lights *19 zeds*

□ Flowers *23 zeds*

Maximum Score Points: 2
Content Domain: Number
Topic Area: Whole Numbers
Cognitive Domain: Reasoning

Exhibit 8 presents the results for the “Decorations” item. More than half the students (51 to 75%) in nearly all of the countries provided the fully correct response. Partial credit (1 point) was awarded for balloons and flowers (48 zeds) or lights, banner, and flowers (47 zeds). Almost three-fourths of the students on average received full or partial credit. On average across countries, a higher percentage of boys answered correctly than girls.

Exhibit 8

School Party Screen 4 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Balloons, Banner, and Lights for 49 zeds)			Percent Partial Credit	
	Overall Country	Girls	Boys	Balloons and Flowers for 48 zeds	Lights, Banner, and Flowers for 47 zeds
England	75 (2.0)	75 (3.6)	75 (2.8)	7 (0.9)	3 (0.9)
Netherlands	71 (2.7)	71 (3.3)	72 (3.8)	8 (1.4)	3 (0.9)
Finland	70 (2.1)	73 (2.5)	68 (3.1)	6 (1.0)	3 (1.1)
Russian Federation	69 (1.9)	66 (2.9)	72 (2.2)	12 (1.3)	2 (0.6)
Denmark	69 (2.7)	66 (3.7)	72 (3.7)	8 (1.8)	2 (0.8)
Sweden	69 (2.4)	70 (3.4)	68 (3.7)	5 (1.4)	4 (1.0)
Norway (5)	69 (2.3)	68 (3.6)	69 (3.2)	5 (1.1)	3 (0.8)
Lithuania	68 (1.9)	71 (3.1)	66 (3.2)	11 (1.4)	3 (0.9)
Singapore	67 (1.8)	66 (2.5)	68 (2.4)	12 (1.1)	6 (0.9)
United States	66 (1.5)	65 (2.0)	66 (2.2)	6 (0.8)	3 (0.6)
Hong Kong SAR	64 (2.7)	60 (3.5)	69 (3.7)	10 (1.6)	6 (1.1)
Croatia	64 (2.9)	61 (3.4)	67 (3.9)	10 (1.4)	3 (1.0)
Czech Republic	63 (2.4)	60 (3.0)	66 (3.4)	10 (1.4)	3 (0.7)
Austria	62 (2.5)	58 (3.6)	66 (3.5)	7 (1.2)	4 (1.0)
Hungary	62 (2.0)	58 (2.9)	65 (3.0)	10 (1.4)	3 (0.7)
Slovak Republic	61 (2.6)	58 (3.6)	65 (3.4)	10 (1.4)	3 (0.9)
Malta	61 (2.0)	60 (2.9)	62 (2.7)	11 (1.4)	4 (0.9)
Spain	60 (2.2)	57 (2.9)	64 (2.8)	10 (1.4)	3 (0.7)
Germany	60 (2.5)	58 (3.7)	62 (3.3)	5 (1.1)	5 (1.2)
Portugal	59 (2.0)	57 (2.7)	60 (2.9)	9 (1.6)	3 (0.7)
Canada	59 (1.6)	53 (2.2)	64 (2.0)	7 (0.8)	4 (0.6)
Korea, Rep. of	56 (2.2)	55 (3.2)	57 (3.5)	8 (1.1)	8 (1.3)
France	53 (2.7)	53 (4.1)	53 (3.2)	7 (1.1)	8 (1.3)
Chile	53 (2.4)	49 (3.5)	55 (3.1)	4 (0.9)	3 (0.7)
Turkey (5)	52 (2.5)	51 (3.5)	53 (3.0)	12 (1.5)	6 (1.1)
Italy	51 (2.3)	45 (3.4)	56 (3.3)	8 (1.3)	1 (0.5)
Chinese Taipei	49 (2.3)	49 (3.8)	49 (3.2)	17 (1.8)	3 (0.8)
United Arab Emirates	41 (1.2)	39 (1.7)	44 (1.4)	8 (0.6)	3 (0.3)
Qatar	34 (2.2)	32 (2.5)	36 (3.0)	8 (1.2)	2 (0.5)
Georgia	34 (2.5)	33 (4.0)	34 (3.4)	10 (1.9)	1 (0.8)
International Average	**60 (0.4)**	**58 (0.6)**	**61 (0.6)** ▲	**9 (0.2)**	**4 (0.2)**
Benchmarking Participants					
Moscow City, Russian Fed.	72 (2.2)	74 (2.9)	69 (3.3)	13 (1.6)	3 (0.8)
Madrid, Spain	61 (2.3)	56 (3.5)	66 (3.3)	10 (1.5)	4 (0.9)
Ontario, Canada	60 (2.7)	54 (4.1)	64 (3.3)	7 (1.4)	3 (1.0)
Quebec, Canada	60 (2.9)	55 (3.7)	65 (3.9)	6 (1.1)	4 (1.0)
Dubai, UAE	58 (1.8)	52 (2.8)	64 (2.5)	9 (1.4)	4 (0.7)
Abu Dhabi, UAE	38 (1.3)	35 (2.3)	40 (2.1)	5 (0.8)	3 (0.6)

▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

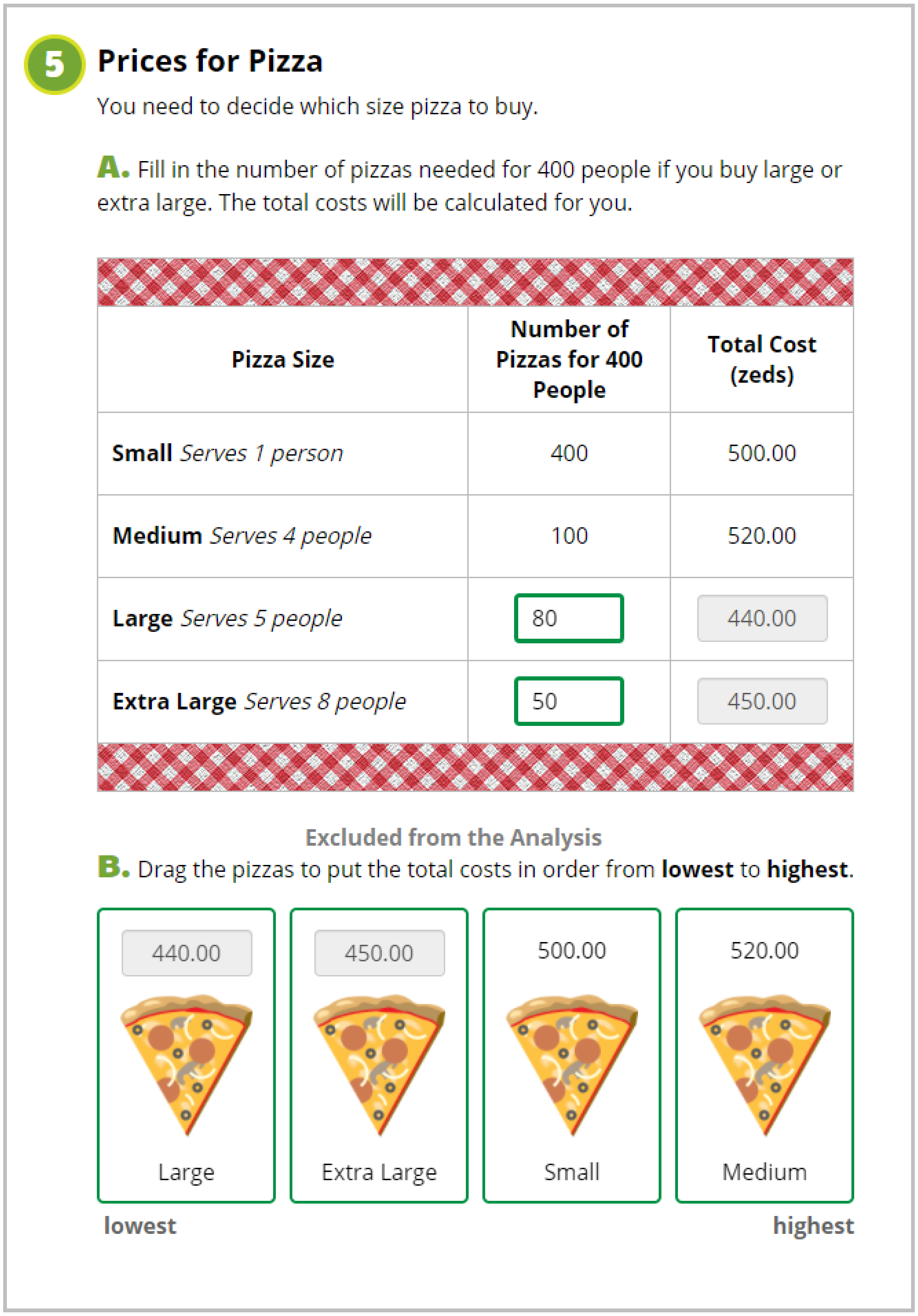
SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 5 – Prices of Pizza

Screen 5 was devoted to deciding what size pizzas to order for the party. The primary stimulus only gave the directive to fill in the number of pizzas needed to serve 400 people for **large** and **extra large** pizzas and that the costs would be calculated for them.

Students needed to understand that the data about the prices of the pizzas were presented in order by the size of the pizzas from smallest to largest. In addition, they needed to comprehend 1) the number of people served by the **small** and the **medium** pizzas, 2) the number of each size needed for 400 people, and 3) the total cost of each size. Beyond that they needed to understand that in lieu of directions, the data for the smaller sizes provided examples of how the students themselves should respond for the two larger sizes. Finally, after all that, it may have been difficult to accept that supplying the answers to the number of people served by each of two larger sizes of pizzas fulfilled the requirements for completing the item. How often does a mathematics test automatically calculate the total cost?

If the fourth grade students managed the reading, they only needed to perform two calculations or do some reasoning to provide 1) the number of **large** pizzas that would be needed to serve 400 people (400 ÷ 5 = 80) and 2) the number of **extra large** pizzas that would be needed to serve 400 people (400 ÷ 8 = 50).



5 Prices for Pizza

You need to decide which size pizza to buy.

A. Fill in the number of pizzas needed for 400 people if you buy large or extra large. The total costs will be calculated for you.

Pizza Size	Number of Pizzas for 400 People	Total Cost (zeds)
Small *Serves 1 person*	400	500.00
Medium *Serves 4 people*	100	520.00
Large *Serves 5 people*	80	440.00
Extra Large *Serves 8 people*	50	450.00

Excluded from the Analysis

B. Drag the pizzas to put the total costs in order from **lowest** to **highest**.

440.00	450.00	500.00	520.00
Large	Extra Large	Small	Medium
lowest			highest

	Item 5A	Item 5B
Maximum Score Points:	2	*Excluded from Scaling and Analysis*
Content Domain:	Number	Number
Topic Area:	Whole Numbers	Whole Numbers
Cognitive Domain:	Applying	Knowing

Exhibit 9 shows the results for Screen 5. It appears that if students understood the type of information they were asked to supply, they were able to answer the number of both sizes of pizza correctly. Unfortunately, however, the majority of students had difficulty. Less than half the fourth grade students in any of the eTIMSS countries answered both questions correctly for full credit (2 points). High-achieving Singapore and Hong Kong SAR managed 44 and 40 percent, respectively. Very few students, 2 to 4 percent on average, answered one but not both of the two sizes correctly. Boys had higher achievement than girls on average across countries.

Exhibit 9

School Party Screen 5 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Both Parts Correct)			Percent Partial Credit (Only 1 Part Correct)	
	Overall Country	Girls	Boys	Number of Large Pizzas Correct	Number of Extra Large Pizzas Correct
Singapore	44 (2.1)	42 (2.7)	46 (2.7)	3 (0.6)	4 (0.7)
Hong Kong SAR	40 (2.7)	35 (3.5)	47 (3.9)	3 (0.8)	4 (0.8)
Russian Federation	39 (2.0)	37 (3.0)	41 (3.0)	2 (0.6)	3 (0.7)
Chinese Taipei	39 (2.7)	37 (3.7)	42 (3.3)	3 (0.8)	3 (0.8)
Korea, Rep. of	34 (2.4)	31 (2.9)	38 (3.7)	1 (0.4)	4 (0.8)
England	30 (2.8)	29 (3.7)	30 (3.3)	4 (0.9)	6 (1.3)
Lithuania	24 (1.9)	25 (2.9)	24 (2.7)	2 (0.6)	3 (0.9)
Germany	20 (2.2)	20 (3.0)	19 (2.5)	3 (1.0)	9 (1.2)
United States	19 (1.4)	16 (1.8)	22 (2.0)	4 (0.9)	3 (0.5)
Georgia	19 (2.3)	13 (2.8)	24 (3.3)	2 (0.9)	3 (0.6)
Austria	19 (1.9)	17 (2.5)	20 (2.8)	1 (0.4)	6 (1.3)
Czech Republic	18 (1.5)	16 (2.1)	20 (2.2)	2 (0.6)	3 (0.7)
Turkey (5)	18 (1.9)	19 (2.6)	17 (2.3)	1 (0.5)	2 (0.6)
Hungary	16 (1.7)	15 (2.1)	17 (2.3)	1 (0.4)	4 (0.7)
Netherlands	15 (1.6)	15 (2.9)	15 (2.4)	1 (0.6)	7 (1.8)
Italy	14 (2.0)	12 (2.4)	16 (2.8)	3 (0.8)	4 (1.0)
Malta	14 (1.7)	12 (2.1)	16 (2.4)	2 (0.8)	3 (0.7)
Croatia	13 (2.2)	17 (3.7)	10 (1.9)	2 (0.6)	3 (0.8)
Portugal	13 (1.4)	12 (2.1)	14 (2.2)	1 (0.4)	3 (0.7)
Denmark	13 (2.2)	12 (2.7)	13 (3.1)	5 (0.9)	8 (1.6)
Norway (5)	11 (1.7)	13 (2.9)	9 (1.8)	5 (1.2)	8 (1.6)
Finland	11 (1.5)	7 (1.4)	15 (2.6)	4 (1.0)	4 (0.7)
Slovak Republic	11 (1.3)	8 (1.6)	13 (2.1)	1 (0.4)	5 (0.9)
Sweden	10 (1.8)	12 (2.9)	9 (1.9)	6 (1.5)	5 (1.3)
United Arab Emirates	10 (0.6)	9 (0.7)	12 (1.0)	2 (0.3)	2 (0.3)
Qatar	9 (1.4)	7 (1.5)	11 (2.3)	1 (0.5)	2 (0.7)
Spain	8 (1.0)	6 (1.4)	9 (1.7)	2 (0.8)	6 (1.2)
Canada	7 (0.9)	6 (1.1)	8 (1.5)	3 (0.5)	4 (0.7)
France	5 (0.8)	4 (0.9)	7 (1.5)	2 (0.5)	4 (1.0)
Chile	4 (0.8)	3 (1.1)	4 (1.3)	2 (0.7)	1 (0.5)
International Average	**18 (0.3)**	**17 (0.5)**	**20 (0.5)** ▲	**2 (0.1)**	**4 (0.2)**
Benchmarking Participants					
Moscow City, Russian Fed.	41 (2.6)	41 (4.0)	41 (3.0)	2 (0.7)	4 (1.1)
Dubai, UAE	18 (1.4)	15 (1.7)	21 (2.3)	2 (0.5)	3 (0.7)
Madrid, Spain	15 (2.2)	15 (3.1)	15 (2.6)	2 (0.7)	3 (1.0)
Ontario, Canada	9 (1.6)	7 (2.0)	11 (2.7)	3 (0.8)	4 (1.1)
Abu Dhabi, UAE	6 (0.8)	5 (1.1)	8 (1.1)	2 (0.4)	2 (0.4)
Quebec, Canada	5 (1.1)	5 (1.3)	6 (1.6)	2 (0.5)	5 (0.9)

▲ Percentage significantly higher than other gender

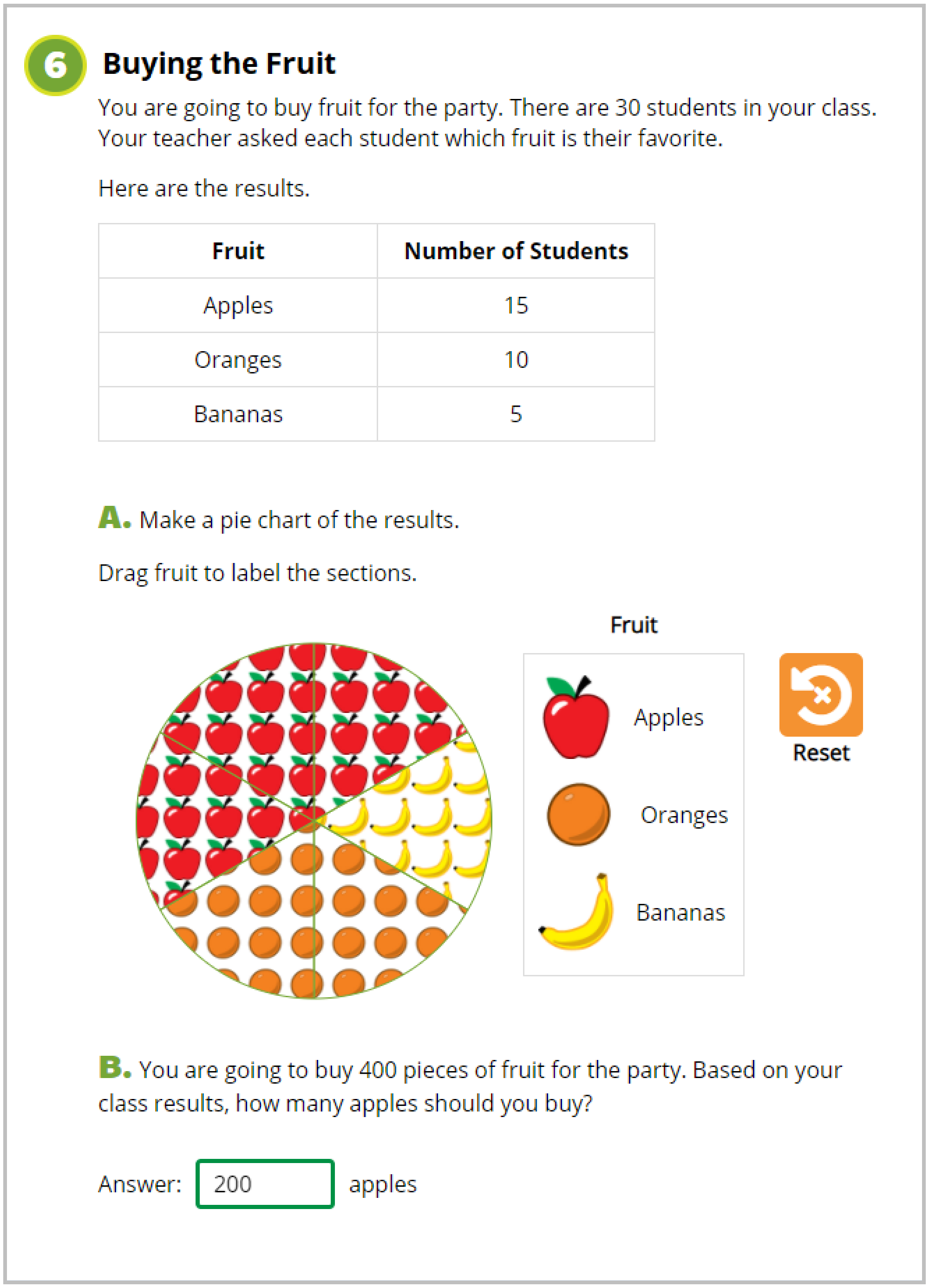
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

5B did not yield meaningful results. The chances of having sensible data about total costs to arrange depended on providing correct answers in 5B. Thus, as a consequence of students' low rates of success on 5B, few students had reasonable data to arrange. Across the countries, 25 percent of the students on average skipped 5B entirely.

Screen 6 – Buying the Fruit

Item 6A asking students to complete a pie chart based on a table of results was among the items with the highest achievement on the *School Party* PSI task. The fourth grade students felt confident about completing the pie chart even this late in the assessment, because nearly all of the students still working on the task attempted this item. They also found it engaging. When a student drags one of the fruits, for example "apples," onto the chart, the section autofills with apples, providing some colorful interactivity. There were 6 sections in the chart, and analysis of the process data found that across the countries students averaged 8 autofills (revising their original answers). Interestingly, some students really liked the autofill feature. The maximum number of autofills was 97 by a student in the Netherlands, a student in Hong Kong SAR had 94, in the United States 92, and Spain 91. Across countries students averaged a maximum of 54 autofills.



6 **Buying the Fruit**

You are going to buy fruit for the party. There are 30 students in your class. Your teacher asked each student which fruit is their favorite.

Here are the results.

Fruit	Number of Students
Apples	15
Oranges	10
Bananas	5

A. Make a pie chart of the results.

Drag fruit to label the sections.

B. You are going to buy 400 pieces of fruit for the party. Based on your class results, how many apples should you buy?

Answer: 200 apples

	Item 6A	Item 6B
Maximum Score Points:	1	1
Content Domain:	Data	Data
Topic Area:	Reading, Interpreting, and Representing	Using Data to Solve Problems
Cognitive Domain:	Applying	Reasoning

On average, 65 percent of the fourth grade eTIMSS students correctly applied the data in the table of class to create a pie chart (see Exhibit 10). Hong Kong SAR performed very well with 90 percent correct, followed by Singapore (84%) and Norway (81%). On average across countries, boys had slightly higher achievement than girls.

Exhibit 10

School Party Screen 6A – Percent Correct Overall and by Gender

Country	Percent Correct (Makes Correct Pie Chart)		
	Overall Country	Girls	Boys
Hong Kong SAR	90 (1.7)	91 (2.1)	89 (2.3)
Singapore	84 (1.5)	82 (2.2)	85 (1.9)
Norway (5)	81 (2.2)	82 (2.9)	80 (3.2)
Korea, Rep. of	81 (1.6)	81 (2.4)	81 (2.2)
Chinese Taipei	77 (1.9)	76 (2.9)	78 (3.0)
Sweden	76 (2.7)	82 (3.0)	71 (3.3)
Russian Federation	76 (1.9)	74 (3.1)	77 (2.1)
Netherlands	75 (2.5)	74 (3.4)	77 (3.7)
Denmark	75 (2.4)	78 (2.8)	72 (3.8)
England	74 (2.0)	73 (3.4)	74 (2.9)
Finland	73 (2.4)	72 (3.3)	74 (2.6)
Canada	73 (1.5)	73 (2.0)	73 (1.8)
Lithuania	73 (2.4)	71 (3.2)	74 (3.2)
United States	64 (1.6)	65 (2.3)	63 (2.2)
Portugal	63 (2.1)	62 (3.3)	65 (2.8)
Austria	63 (2.6)	60 (3.1)	66 (3.9)
Germany	63 (2.4)	64 (3.7)	62 (3.1)
Czech Republic	63 (2.3)	63 (3.3)	63 (3.4)
Slovak Republic	61 (2.3)	55 (3.7)	66 (2.9)
Spain	61 (2.0)	56 (3.1)	65 (3.0)
Turkey (5)	61 (2.3)	59 (3.2)	63 (3.5)
Italy	60 (3.0)	57 (4.1)	64 (4.1)
Hungary	60 (2.2)	59 (3.0)	60 (3.2)
Malta	56 (2.3)	53 (3.0)	59 (3.4)
Croatia	55 (3.1)	55 (4.4)	54 (4.4)
United Arab Emirates	45 (1.1)	44 (1.6)	45 (1.3)
France	44 (2.7)	39 (3.5)	49 (3.3)
Qatar	42 (2.4)	43 (3.1)	40 (3.4)
Georgia	39 (2.8)	36 (4.4)	42 (3.7)
Chile	37 (2.6)	34 (3.3)	39 (3.2)
International Average	**65 (0.4)**	**64 (0.6)**	**66 (0.6)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	86 (1.6)	86 (2.1)	86 (2.3)
Ontario, Canada	79 (2.5)	78 (3.3)	80 (3.0)
Quebec, Canada	72 (2.4)	75 (2.9)	68 (4.0)
Madrid, Spain	67 (2.0)	62 (3.4)	72 (2.9)
Dubai, UAE	64 (1.9)	63 (2.6)	64 (2.3)
Abu Dhabi, UAE	35 (1.5)	34 (2.4)	36 (2.1)

Percentage significantly higher than other gender ▲

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

In contrast to the high performance on the pie chart, 6B was one of the most difficult items in this PSI task. Students were asked to use the class results where 15 out of 30 students voted for apples, to decide how many of the 400 pieces of fruit being bought for the party should be apples. Eighteen percent of the students on average skipped over this problem that required simple proportional reasoning based on one-half (or several steps of calculations).

Exhibit 11 shows the low percentages of correct responses provided by students in eTIMSS countries, with only 14 percent of the students on average providing a correct response. Korea with 28 percent was the highest performing country, but the rest of the eTIMSS countries had 25 percent correct or less (a number of countries had less than 20 percent or even 10 percent correct). Because $\frac{3}{6}$ and $\frac{1}{2}$ were relatively common among the incorrect responses, some students understood that half the pieces of fruit should be apples but did not convert that fraction to determine that the party planners should buy 200 apples for the party.

In contrast to the small gender differences in the relatively high levels of success shown for making the pie chart, Part B revealed considerable gender differences across countries on average favoring boys, although both boys and girls had relatively low percentages of success in using proportional reasoning or calculations to determine the number of apples to buy.

Exhibit 11

School Party Screen 6B – Percent Correct Overall and by Gender

Country	Percent Correct (200 Apples)		
	Overall Country	Girls	Boys
Korea, Rep. of	28 (2.3)	23 (2.9)	33 (3.1)
England	25 (2.1)	17 (2.7)	31 (3.0)
Sweden	24 (2.6)	19 (3.3)	30 (3.9)
Hong Kong SAR	24 (2.4)	16 (3.0)	32 (3.3)
Finland	23 (1.8)	23 (2.8)	23 (2.7)
Norway (5)	21 (2.3)	23 (3.6)	20 (2.8)
Denmark	21 (2.1)	19 (3.3)	24 (2.9)
Netherlands	20 (2.0)	19 (2.7)	21 (2.9)
Singapore	19 (1.4)	17 (2.0)	20 (2.2)
Chinese Taipei	17 (2.0)	12 (1.9)	23 (3.0)
Russian Federation	16 (1.5)	13 (2.3)	19 (2.4)
Slovak Republic	15 (2.0)	9 (2.1)	20 (3.3)
United States	14 (1.2)	11 (1.7)	17 (1.8)
Lithuania	14 (1.8)	9 (1.7)	18 (3.0)
Canada	13 (1.2)	12 (1.8)	15 (1.5)
Czech Republic	13 (1.7)	12 (1.8)	15 (2.7)
Germany	12 (1.6)	9 (2.3)	14 (2.4)
Hungary	12 (1.6)	11 (2.3)	12 (2.2)
Italy	12 (1.9)	11 (2.4)	12 (2.4)
Portugal	11 (1.4)	8 (1.7)	12 (2.2)
Spain	10 (1.3)	5 (1.2)	14 (2.1)
United Arab Emirates	10 (0.6)	8 (0.7)	11 (0.9)
Austria	9 (1.3)	8 (2.2)	10 (1.6)
Turkey (5)	9 (1.3)	8 (2.0)	9 (1.8)
France	9 (1.5)	7 (1.8)	11 (2.1)
Malta	7 (1.3)	3 (1.1)	11 (2.1)
Croatia	7 (1.9)	9 (3.8)	5 (1.5)
Chile	6 (1.2)	5 (1.6)	7 (1.6)
Georgia	6 (1.2)	5 (1.7)	7 (1.9)
Qatar	5 (1.2)	5 (1.8)	5 (1.5)
International Average	**14 (0.3)**	**12 (0.4)**	**17 (0.4)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	17 (1.8)	16 (2.5)	18 (2.9)
Quebec, Canada	15 (1.9)	11 (2.5)	20 (2.7)
Dubai, UAE	15 (1.4)	11 (1.6)	18 (2.5)
Ontario, Canada	13 (2.0)	13 (3.4)	13 (2.0)
Madrid, Spain	10 (1.6)	7 (2.1)	12 (2.1)
Abu Dhabi, UAE	6 (0.7)	5 (0.9)	8 (1.2)

Percentage significantly higher than other gender ▲

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

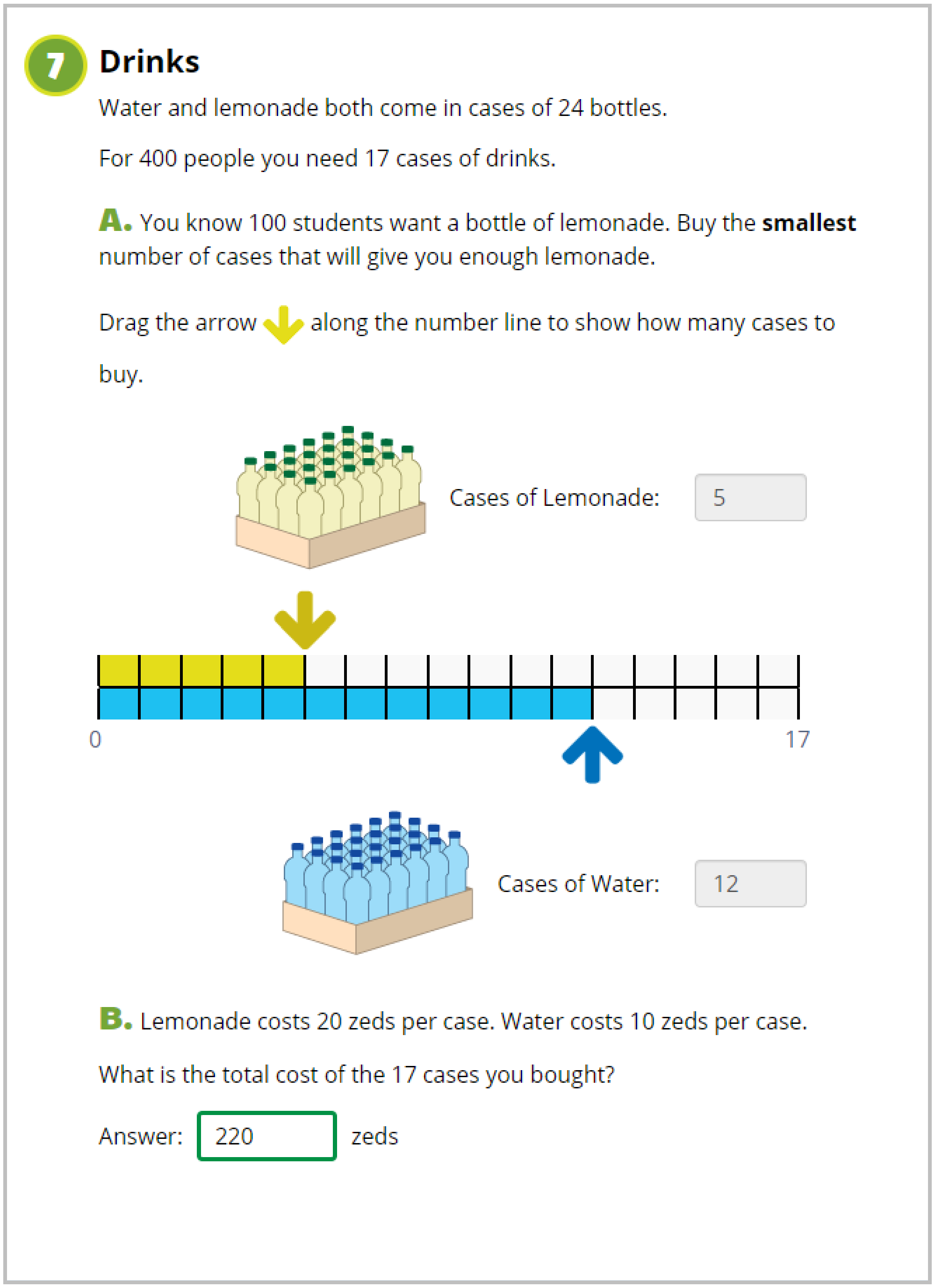
SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 7 – Drinks—Lemonade or Water

By the beginning of Screen 7 with four items left in this PSI (2 items on this screen and 2 on Screen 8), the percent of not-reached students increased to 11 percent. Of the students that entered Screen 7 with assessment time remaining, 15 percent still omitted 7A.

Screen 7's introductory information included: 1) the cases of drinks contain 24 bottles, and 2) 400 people will need 17 cases of drinks. 7A involved determining the smallest number of cases that would provide enough lemonade for 100 people. To provide their answers, the fourth grade students were asked to use a slider tool to drag a yellow arrow along a number line with 17 unit marks to indicate the 17 cases of drinks (but without any numbers). The slider tool was intended to be engaging and reduce the number of calculations necessary in 7B, where students were asked to determine the total cost of the drinks. However, the purpose of the slider tool may not have been clear to the students.

To solve 7A, students were expected to determine (one way or another) that 4 cases of 24 bottles of lemonade would not be enough lemonade for 100 people (because it is 4 bottles short). Then, if they had been using the typical eTIMSS response mode for a numerical response, the students would have used their number pad to enter 5 (which would have been recorded in the answer box). However, in this particular item students were asked to "drag the yellow arrow along the number line to show how many cases to buy." The intention was for them to move the arrow 5 units, which automatically would make 5 appear in the box labelled "Cases of Lemonade" and 12 appear in the box labelled "Cases of Water."



7 Drinks

Water and lemonade both come in cases of 24 bottles.

For 400 people you need 17 cases of drinks.

A. You know 100 students want a bottle of lemonade. Buy the **smallest** number of cases that will give you enough lemonade.

Drag the arrow ↓ along the number line to show how many cases to buy.

Cases of Lemonade: 5

Cases of Water: 12

B. Lemonade costs 20 zeds per case. Water costs 10 zeds per case.

What is the total cost of the 17 cases you bought?

Answer: 220 zeds

	Item 7A	Item 7B
Maximum Score Points:	1	1
Content Domain:	Number	Number
Topic Area:	Whole Numbers	Whole Numbers
Cognitive Domain:	Reasoning	Reasoning

However, only 15 percent of the students on average across countries moved the arrow 5 units (Exhibit 12). The highest achievement, 25 to 29 percent correct responses, was in Hong Kong SAR, Norway, Singapore, Korea, and Chinese Taipei. Further analysis of response data indicated that some students (8% on average) moved the arrow four units (causing 4 to appear by the cases of lemonade and 13 to appear by the cases of water). It may be that the students moving the arrow four units just did not round up to 5 cases for some reason. On average across countries, boys had higher percentages of correct responses than girls.

Exhibit 12

School Party Screen 7A – Percent Correct Overall and by Gender

Country	Percent Correct (5 Cases of Lemonade)		
	Overall Country	Girls	Boys
Hong Kong SAR	29 (2.8)	24 (3.5)	35 (3.7)
Norway (5)	29 (2.3)	29 (3.7)	29 (3.8)
Singapore	26 (1.9)	25 (2.7)	26 (2.4)
Korea, Rep. of	25 (2.3)	23 (2.8)	27 (3.2)
Chinese Taipei	25 (2.3)	25 (3.0)	25 (3.4)
England	21 (2.1)	19 (3.2)	23 (2.7)
Denmark	19 (2.1)	18 (2.7)	19 (3.3)
Netherlands	17 (2.1)	13 (2.4)	21 (2.7)
Lithuania	17 (1.8)	17 (2.6)	16 (2.5)
Russian Federation	16 (1.6)	16 (2.0)	16 (2.7)
Finland	15 (1.6)	11 (1.8)	19 (2.5)
Sweden	15 (1.5)	15 (2.3)	15 (2.5)
Czech Republic	14 (1.9)	11 (1.7)	16 (3.3)
Germany	14 (1.6)	11 (2.5)	16 (2.2)
Canada	13 (1.3)	13 (1.9)	14 (1.6)
United States	13 (0.9)	12 (1.5)	15 (1.6)
Italy	13 (1.9)	9 (2.1)	16 (2.7)
Hungary	13 (1.5)	12 (2.1)	13 (2.3)
Austria	13 (1.7)	10 (2.2)	14 (2.5)
Slovak Republic	12 (1.6)	9 (2.2)	15 (2.5)
Malta	11 (1.4)	9 (1.8)	13 (2.1)
Turkey (5)	10 (1.5)	9 (2.1)	10 (2.0)
Croatia	10 (1.4)	9 (2.1)	10 (1.9)
Portugal	9 (1.3)	10 (1.9)	8 (1.9)
Georgia	9 (1.8)	7 (2.7)	10 (2.3)
Spain	9 (1.5)	5 (1.1)	12 (2.3)
France	8 (1.3)	7 (1.5)	9 (1.8)
United Arab Emirates	8 (0.5)	7 (0.8)	9 (0.7)
Qatar	7 (1.3)	9 (2.3)	5 (1.5)
Chile	6 (1.4)	4 (1.5)	8 (1.8)
International Average	**15 (0.3)**	**13 (0.4)**	**16 (0.5)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	20 (1.9)	21 (2.3)	18 (2.7)
Quebec, Canada	15 (2.6)	14 (3.8)	15 (2.8)
Ontario, Canada	14 (2.1)	14 (2.9)	13 (2.6)
Dubai, UAE	12 (1.1)	10 (1.5)	14 (1.7)
Madrid, Spain	12 (1.6)	13 (2.0)	11 (2.1)
Abu Dhabi, UAE	4 (0.7)	3 (0.7)	6 (1.2)

Percentage significantly higher than other gender ▲

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Based on a cursory look at the process data, it appears that the fourth grade students did not actually understand that the slider tool was the way to submit their answers to the item. Most moved the slider tool, so maybe they found it interesting. Yet 71 percent of students on average; that is, most of the students—except the 15 percent responding correctly and the 15 percent that omitted the item—left the arrow at some spot other than 5. Consequently, their Part A did not show the correct data about the cases of lemonade and cases of water, but every possible combination of incorrect data. As an important consideration in developing innovative response options, no matter how engaging, if the students do not understand the purpose of the tools or how to us them, then the tools may be more of a distraction than anything else.

Do not use items that depend on correct answers to previous items. 7B, isolated from dependency on where the slider tool was expected to be in 7A, could have been straightforward. Students should have been given new conditions for cost with new values that would have made the calculations in 7B relatively easy like was originally intended. Based on the current 7A and 7B, only the students with a correct answer to Part A have the correct multipliers for the cases of lemonade and cases of water necessary to calculate a correct answer to 7B.

Exhibit 13 containing the results for 7B shows only 8 percent of the students on average managed the correct calculations for the total cost of the drinks. Even with correct values, finding the solution involved three steps. That is: 5 cases × 20 zeds = 100 zeds for the lemonade and 12 cases × 10 zeds = 120 zeds for the water. Then, 100 zeds + 120 zeds = 220 zeds. The scoring guide also included tracking for the students whose calculations for the total cost matched up with an incorrect number of cases of lemonade (any number other than 5). This elevated the 8 percent correct on average by another 15 percent, with nearly reaching one-fourth of the students providing correct multiplication and addition in 7B.

Exhibit 13

School Party Screen 7B – Percent Correct Overall and by Gender

Country	Percent Correct using Correct Number of Cases of Lemonade			Percent Correct using Incorrect Number of Cases of Lemonade
	Overall Country	Girls	Boys	
Hong Kong SAR	21 (2.3)	16 (2.8)	26 (3.4)	27 (2.8)
Singapore	18 (1.7)	18 (2.3)	18 (2.2)	25 (1.7)
Korea, Rep. of	16 (2.0)	13 (2.3)	19 (2.8)	22 (1.8)
Norway (5)	15 (1.9)	13 (2.4)	17 (3.2)	16 (1.9)
Chinese Taipei	14 (1.7)	12 (2.1)	16 (2.5)	17 (1.7)
England	13 (2.0)	9 (2.8)	15 (2.8)	21 (2.4)
Denmark	11 (1.8)	10 (2.1)	12 (2.8)	20 (2.7)
Russian Federation	10 (1.5)	9 (1.9)	11 (2.5)	20 (1.8)
Netherlands	10 (1.8)	8 (2.4)	12 (2.2)	16 (1.8)
United States	9 (0.8)	7 (1.3)	10 (1.3)	19 (1.1)
Lithuania	8 (1.0)	8 (1.9)	7 (1.2)	12 (1.9)
Germany	7 (1.3)	7 (2.1)	7 (1.6)	17 (1.9)
Slovak Republic	7 (1.1)	4 (1.4)	9 (1.9)	12 (1.5)
Czech Republic	6 (1.1)	6 (1.4)	7 (1.6)	15 (1.6)
Canada	6 (0.8)	6 (1.2)	7 (1.0)	15 (1.4)
Austria	6 (1.3)	5 (1.3)	8 (2.0)	17 (2.0)
Finland	6 (1.1)	5 (1.2)	7 (1.6)	12 (1.2)
Italy	6 (1.2)	4 (1.3)	8 (2.0)	8 (1.3)
Sweden	5 (0.9)	6 (1.2)	5 (1.5)	9 (1.6)
Malta	5 (1.0)	4 (1.3)	6 (1.5)	12 (1.5)
Turkey (5)	5 (1.0)	4 (1.2)	6 (1.5)	13 (1.7)
Hungary	5 (0.9)	4 (1.2)	5 (1.4)	17 (1.8)
France	4 (0.8)	3 (0.7)	6 (1.5)	12 (1.7)
Portugal	4 (1.1)	3 (1.1)	5 (1.6)	16 (2.0)
United Arab Emirates	4 (0.4)	3 (0.5)	4 (0.6)	8 (0.5)
Spain	4 (0.7)	2 (0.7)	5 (1.1)	15 (1.4)
Croatia	3 (0.9)	1 (0.5)	6 (1.5)	12 (2.4)
Georgia	3 (1.1)	2 (1.4)	5 (1.7)	9 (1.6)
Chile	2 (0.7)	3 (1.1)	2 (0.9)	7 (1.2)
Qatar	2 (0.6)	3 (1.0)	2 (1.0)	5 (0.8)
International Average	**8 (0.2)**	**7 (0.3)**	**9 (0.4)** ▲	**15 (0.3)**
Benchmarking Participants				
Moscow City, Russian Fed.	11 (1.4)	11 (1.9)	11 (2.2)	25 (1.8)
Dubai, UAE	8 (1.0)	7 (1.4)	8 (1.3)	13 (1.2)
Quebec, Canada	7 (1.4)	6 (1.7)	7 (2.2)	16 (2.0)
Ontario, Canada	6 (1.3)	6 (2.3)	6 (1.4)	15 (2.0)
Madrid, Spain	5 (1.4)	5 (1.5)	6 (1.9)	16 (2.2)
Abu Dhabi, UAE	2 (0.5)	1 (0.3)	3 (0.9)	5 (0.6)

▲ Percentage significantly higher than other gender

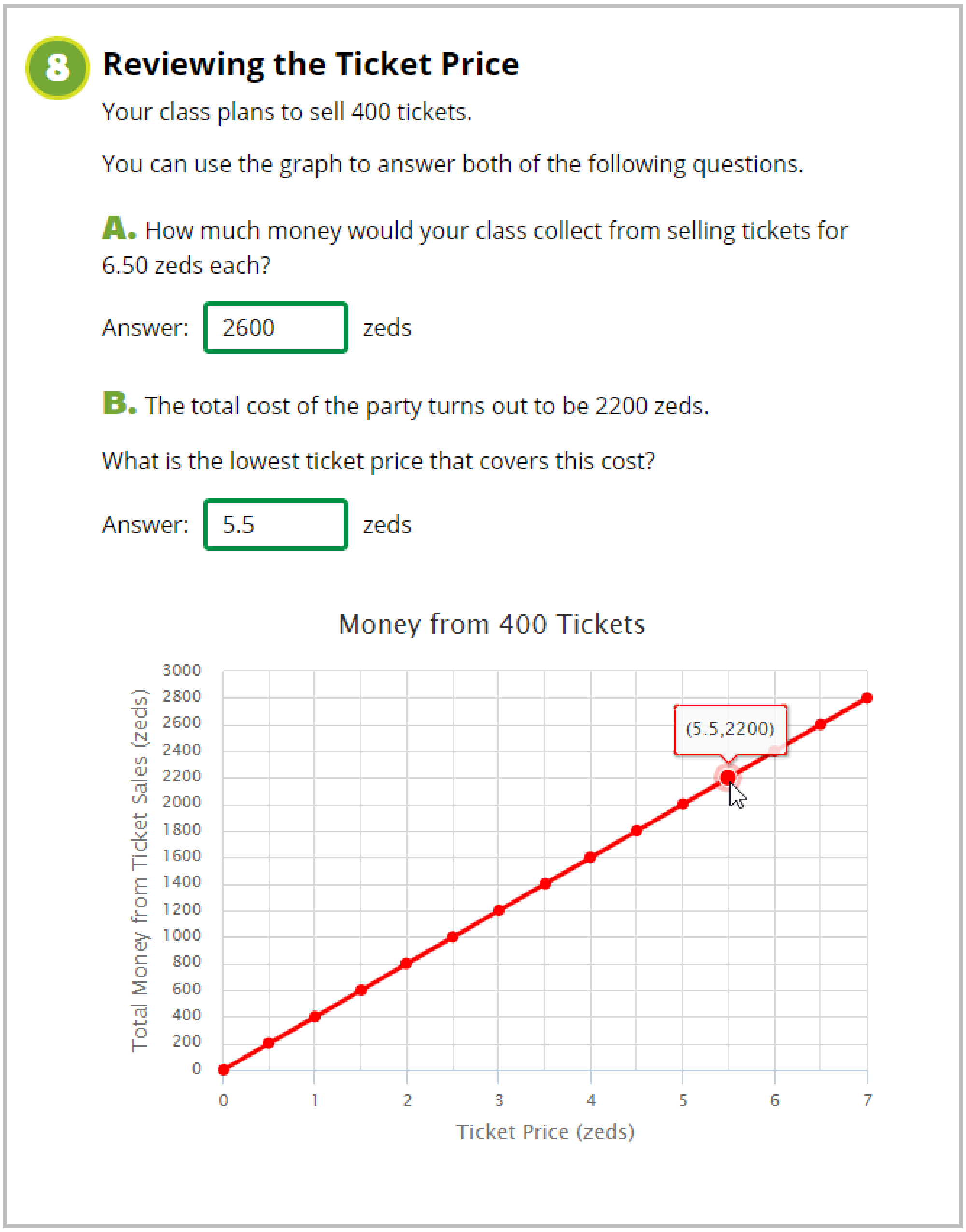
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 8 – Reviewing the Ticket Price

The last screen in the task included two items, each asking students to read a line graph showing the relationship between ticket price for 400 people on the *x*-axis and total cost of the party on the *y*-axis. On average, 16 percent of the students did not reach 8A and 20 percent did not reach 8B.

8A asked students how much money the class would make by selling 400 tickets for 6.50 zeds each.



8 Reviewing the Ticket Price

Your class plans to sell 400 tickets.

You can use the graph to answer both of the following questions.

A. How much money would your class collect from selling tickets for 6.50 zeds each?

Answer: 2600 zeds

B. The total cost of the party turns out to be 2200 zeds.

What is the lowest ticket price that covers this cost?

Answer: 5.5 zeds

	Item 8A	Item 8B
Maximum Score Points:	1	1
Content Domain:	Data	Data
Topic Area:	Reading, Interpreting, and Representing	Using Data to Solve Problems
Cognitive Domain:	Knowing	Reasoning

Exhibit 14 containing the results reveals that only 15 percent of the students on average were able to find 6.5 on the *x*-axis (indicated by a gridline between 6 and 7 zeds) and then read across two gridlines down to find 2600. The highest performance was 34 percent correct in Chinese Taipei, with less than one-fourth of the students responding correctly in nearly all the countries. On average across countries, a higher percentage of boys than girls responded correctly.

Exhibit 14

School Party Screen 8A – Percent Correct Overall and by Gender

Country	Percent Correct (2600 zeds)		
	Overall Country	Girls	Boys
Chinese Taipei	34 (2.4)	34 (3.3)	34 (3.1)
Russian Federation	26 (2.1)	20 (2.9)	32 (2.8)
Netherlands	25 (2.3)	20 (3.2)	29 (3.7)
Norway (5)	25 (3.0)	21 (4.0)	29 (4.1)
Singapore	24 (1.9)	23 (2.4)	25 (2.5)
Korea, Rep. of	23 (2.2)	19 (2.7)	26 (3.3)
Hong Kong SAR	21 (2.6)	15 (2.7)	29 (3.8)
England	21 (2.4)	15 (3.2)	27 (3.5)
Sweden	20 (2.6)	17 (3.4)	23 (3.7)
Denmark	18 (2.7)	13 (2.7)	23 (4.1)
Portugal	18 (2.2)	18 (2.8)	17 (2.8)
Georgia	17 (2.4)	7 (2.5)	24 (3.6)
Lithuania	15 (1.8)	12 (2.7)	18 (2.7)
Italy	15 (1.6)	11 (2.7)	18 (2.3)
Finland	14 (1.5)	11 (2.0)	16 (2.2)
Hungary	14 (1.7)	10 (1.9)	17 (2.6)
Germany	13 (1.7)	8 (2.3)	18 (2.8)
Turkey (5)	13 (1.8)	12 (2.7)	14 (2.2)
United Arab Emirates	13 (0.7)	11 (1.0)	14 (0.9)
Slovak Republic	13 (1.6)	11 (2.5)	14 (2.5)
Czech Republic	12 (1.7)	11 (2.3)	13 (2.4)
Austria	11 (1.6)	9 (2.1)	13 (2.3)
Spain	10 (1.4)	8 (1.8)	12 (2.1)
United States	10 (1.0)	9 (1.3)	12 (1.5)
Croatia	10 (2.1)	6 (3.7)	12 (2.5)
Canada	8 (0.9)	5 (1.0)	10 (1.6)
Malta	6 (1.1)	4 (1.2)	8 (1.7)
France	5 (1.2)	3 (1.0)	8 (1.9)
Qatar	5 (1.1)	3 (1.2)	6 (1.5)
Chile	4 (1.0)	5 (1.7)	3 (1.1)
International Average	**15 (0.3)**	**12 (0.5)**	**18 (0.5)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	25 (2.2)	22 (3.0)	28 (3.2)
Dubai, UAE	17 (1.5)	16 (1.9)	18 (1.9)
Madrid, Spain	11 (1.6)	7 (1.7)	15 (2.9)
Quebec, Canada	9 (1.7)	6 (2.0)	13 (3.1)
Abu Dhabi, UAE	9 (0.9)	6 (1.1)	12 (1.6)
Ontario, Canada	7 (1.5)	3 (1.4)	10 (2.4)

Percentage significantly higher than other gender ▲

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

In Part B, students were asked to find the lowest ticket price that would cover the cost, if the party had a total cost of 2200. Here, students needed to find 2200 on the *y*-axis and then read down to *x*-axis. The ticket price would need to be at least 5.50 zeds (5.5 is on the gridline, but between 5 and 6 zeds).

Exhibit 15 shows the percentages of correct responses for the eTIMSS countries. Achievement on Part B was similar to Part A, except even a little lower—11 percent correct on average. The highest performance ranged from 20 to 22 percent correct (Hong Kong SAR, England, and the Russian Federation). Boys had higher percentages of correct responses than girls on average across countries.

Exhibit 15

School Party Screen 8B – Percent Correct Overall and by Gender

Country	Percent Correct (5.5 zeds)		
	Overall Country	Girls	Boys
Russian Federation	22 (1.9)	22 (2.4)	22 (3.2)
England	21 (2.7)	15 (3.2)	26 (3.8)
Hong Kong SAR	20 (2.5)	17 (2.8)	23 (3.5)
Singapore	19 (1.6)	14 (1.8)	23 (2.3)
Norway (5)	18 (2.3)	13 (3.1)	23 (3.6)
Finland	16 (2.1)	14 (2.3)	19 (2.8)
Denmark	14 (2.1)	11 (2.7)	18 (3.5)
Korea, Rep. of	14 (1.9)	11 (2.1)	18 (2.8)
Netherlands	14 (1.8)	9 (2.3)	18 (2.8)
Chinese Taipei	13 (1.8)	13 (2.1)	14 (2.6)
Portugal	13 (2.0)	11 (2.1)	15 (3.0)
Sweden	12 (2.1)	8 (2.1)	15 (3.2)
Lithuania	12 (1.7)	10 (2.1)	14 (2.5)
Czech Republic	11 (1.8)	11 (2.6)	11 (2.3)
Turkey (5)	10 (1.5)	8 (1.9)	11 (2.0)
Georgia	10 (1.9)	3 (0.9)	15 (3.0)
Germany	10 (1.5)	6 (1.6)	13 (2.6)
Slovak Republic	9 (1.4)	8 (2.0)	10 (2.2)
United Arab Emirates	9 (0.5)	7 (0.6)	11 (0.9)
Italy	9 (1.6)	4 (1.6)	13 (2.6)
United States	7 (0.8)	7 (1.2)	8 (1.1)
Croatia	7 (1.9)	7 (4.4)	7 (2.6)
Hungary	6 (1.1)	5 (1.4)	8 (1.7)
Spain	6 (1.2)	3 (1.1)	9 (1.9)
France	6 (1.4)	5 (1.3)	7 (2.1)
Canada	5 (1.0)	3 (0.8)	7 (1.5)
Austria	4 (1.2)	3 (1.7)	5 (1.6)
Malta	4 (0.9)	2 (1.0)	5 (1.4)
Qatar	4 (1.2)	4 (1.5)	3 (1.2)
Chile	2 (0.7)	3 (1.2)	2 (0.8)
International Average	**11 (0.3)**	**9 (0.4)**	**13 (0.5)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	22 (2.2)	24 (3.3)	21 (2.7)
Dubai, UAE	14 (1.2)	12 (1.7)	16 (1.8)
Madrid, Spain	8 (1.5)	5 (1.8)	11 (2.3)
Ontario, Canada	5 (1.8)	2 (1.3)	8 (2.8)
Abu Dhabi, UAE	5 (0.6)	3 (0.7)	7 (1.2)
Quebec, Canada	4 (1.0)	3 (1.3)	6 (1.8)

Percentage significantly higher than other gender ▲

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 9 – Party

9 Party

Thank you for helping to plan the party!

Conclusions and Reflections

The scenario underlying the *School Party* PSI task generally was appropriate for the fourth grade students, and it was colorful with a number of interactive features to keep students engaged and motivated. However, taken all together, the task may have been a little too ambitious, resulting in unusually high levels of non-response (see Appendix A) and lower achievement than expected on many of the items.

- *School Party* was a full-length PSI task administered together with other difficult PSI items in a 36-minute session, such that 21 percent of students stopped responding before the end and 14 percent ran out of time. Future fourth grade PSI tasks should only be half as long, mini-PSI tasks, and should be administered in sessions together with regular TIMSS items.
- Interactivity can be very successful in encouraging high performance, as for example, the autofilling pie chart on Screen 6. However, it can create a barrier to accomplishing the task if students cannot recognize what to do, as with the slider for the bottles of lemonade and water on Screen 7.
- While working to realize the potential benefits that technology has to offer, it is crucial to keep in mind the basic principles: avoid item dependency, minimize reading load, and ensure that students know how to input their responses to the items.

CHAPTER 2

Science Grade 4

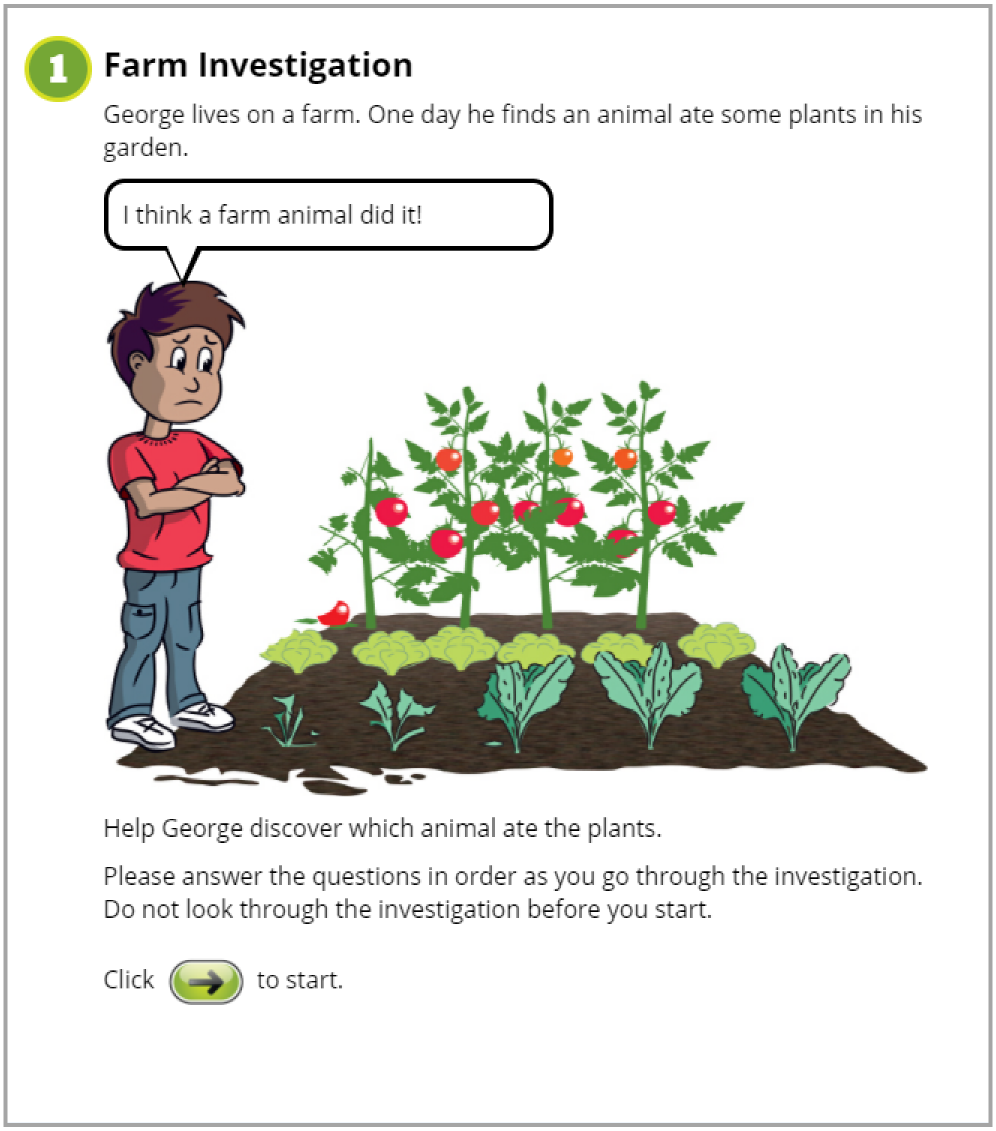
Farm Investigation

About the Task

The *Farm Investigation* Problem Solving and Inquiry (PSI) task, although situated in a life science content domain, was intended primarily to assess fourth grade students' knowledge and competencies in the practices of scientific inquiry. As described in the TIMSS 2019 Science Assessment Framework, five practices fundamental to scientific inquiry are represented in TIMSS 2019: asking questions based on observations, generating evidence, working with data, answering the research question, and making an argument from evidence. As students work through the *Farm Investigation* task to help George discover which animals ate the plants in his garden, they engage with activities that involve these practices.

Screen 1 – Farm Investigation

This screen introduces George and his farm, and his hypothesis that it was a farm animal that ate his plants. Students were requested to answer the questions in order as they worked through the task, and not to look through the investigation before starting.



1 **Farm Investigation**

George lives on a farm. One day he finds an animal ate some plants in his garden.

I think a farm animal did it!

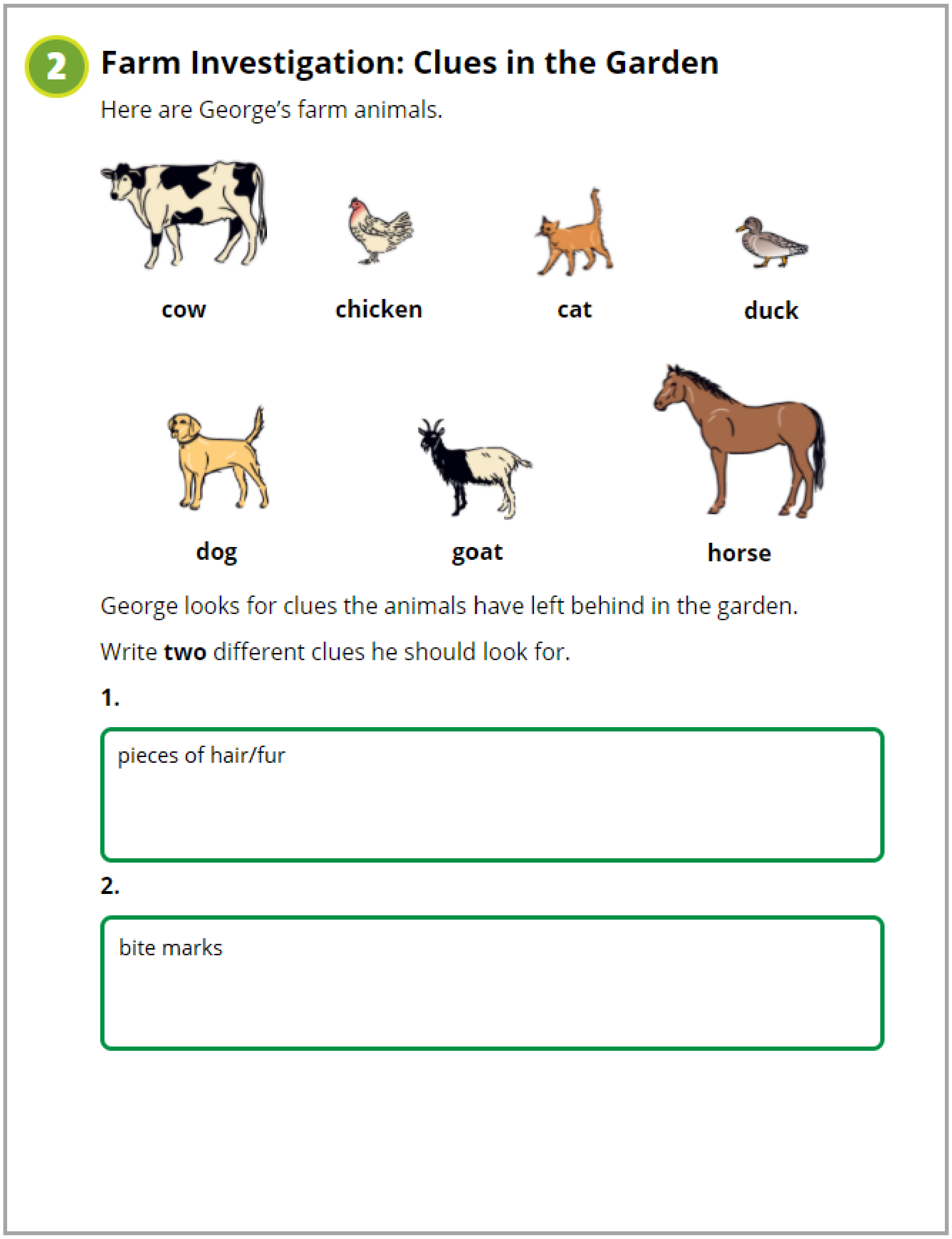
Help George discover which animal ate the plants.

Please answer the questions in order as you go through the investigation. Do not look through the investigation before you start.

Click ➔ to start.

Screen 2 – Clues in the Garden

Screen 2 shows all the animals on the farm (the list of possible “suspects”) and asks the students to suggest two different clues that George should look for to help him decide which animal ate his plants. This question requires some basic knowledge of attributes/characteristics of the farm animals shown, and expects the students to think about what evidence George would need to advance his investigation. For full credit (2 points), students provide two of the following clues: hair/fur, footprints/tracks, poop/scat/excrement, bite marks, eggs, feathers, and which plants were eaten.



2 Farm Investigation: Clues in the Garden

Here are George's farm animals.

cow

chicken

cat

duck

dog

goat

horse

George looks for clues the animals have left behind in the garden.

Write **two** different clues he should look for.

1.

pieces of hair/fur

2.

bite marks

Maximum Score Points: 2
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Applying
Science Practice: Generating Evidence

Exhibit 16 presents the percentage of students in each country, overall and by gender, that earned full credit on the item by providing two acceptable clues. The percentage earning partial credit by providing just one clue also is shown. Students generally found this item to be challenging, with just 25 percent earning full credit on average internationally, and 21 percent earning partial credit. Average performance was highest in Sweden, Hungary, Finland, and England, where 40 percent or more of the students achieved full credit. Girls performed a little better than boys, with 27 percent achieving full credit on average, compared with 24 percent for boys.

Exhibit 16

Farm Investigation Screen 2 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Gives 2 Correct Clues)			Percent Partial Credit (Gives only 1 Correct Clue)
	Overall Country	Girls	Boys	
Sweden	45 (2.7)	47 (3.9)	43 (3.5)	22 (2.0)
Hungary	42 (2.2)	46 (3.5)	38 (3.2)	26 (2.1)
Finland	41 (2.2)	41 (3.7)	41 (2.9)	28 (1.7)
England	40 (2.2)	44 (3.5)	35 (3.1)	27 (2.1)
Denmark	39 (2.3)	43 (3.3)	35 (3.6)	21 (2.0)
Russian Federation	34 (2.0)	33 (2.5)	35 (2.8)	22 (1.6)
Austria	32 (2.4)	34 (3.0)	31 (3.5)	17 (1.7)
Canada	32 (1.8)	35 (2.4)	28 (2.3)	29 (1.8)
United States	31 (1.4)	32 (1.9)	31 (1.8)	22 (1.4)
Czech Republic	31 (2.3)	33 (3.3)	27 (2.8)	25 (1.7)
Singapore	29 (1.7)	33 (2.3)	24 (2.2)	30 (1.7)
Turkey (5)	28 (1.9)	28 (2.8)	28 (2.8)	25 (1.7)
Korea, Rep. of	27 (1.9)	27 (2.8)	27 (2.6)	36 (2.1)
Slovak Republic	27 (2.2)	28 (2.7)	25 (3.4)	26 (2.1)
Malta	26 (2.0)	27 (2.7)	24 (2.7)	16 (1.9)
Italy	26 (2.3)	25 (2.7)	26 (3.5)	22 (2.3)
Lithuania	25 (2.0)	26 (2.7)	24 (3.1)	24 (1.9)
Croatia	25 (2.1)	29 (3.1)	21 (2.8)	16 (1.5)
Spain	24 (1.5)	23 (2.1)	24 (2.6)	23 (1.9)
France	22 (2.1)	24 (2.9)	20 (2.8)	20 (1.7)
Portugal	20 (1.9)	23 (2.6)	18 (2.5)	25 (1.9)
Chile	16 (1.7)	18 (2.4)	15 (2.0)	14 (1.6)
Netherlands	15 (2.1)	13 (2.5)	16 (2.9)	17 (2.0)
United Arab Emirates	14 (0.7)	15 (0.9)	13 (0.9)	15 (0.6)
Germany	14 (1.8)	13 (2.5)	14 (2.4)	10 (1.5)
Georgia	13 (1.7)	12 (2.3)	14 (2.4)	13 (2.2)
Hong Kong SAR	13 (2.2)	12 (1.7)	14 (3.9)	14 (1.8)
Norway (5)	13 (1.6)	16 (2.7)	9 (2.1)	9 (1.4)
Chinese Taipei	12 (1.4)	12 (1.9)	12 (2.3)	21 (2.3)
Qatar	10 (1.1)	12 (1.9)	8 (1.6)	15 (1.6)
International Average	**25 (0.4)**	**27 (0.5)** ▲	**24 (0.5)**	**21 (0.3)**
Benchmarking Participants				
Ontario, Canada	32 (2.9)	37 (4.1)	28 (3.7)	30 (3.1)
Moscow City, Russian Fed.	31 (2.0)	32 (3.3)	29 (3.2)	22 (2.0)
Quebec, Canada	28 (2.9)	30 (4.0)	24 (3.7)	30 (2.4)
Madrid, Spain	27 (2.3)	27 (2.7)	27 (3.1)	22 (2.2)
Dubai, UAE	21 (1.3)	24 (2.0)	19 (1.9)	26 (1.6)
Abu Dhabi, UAE	9 (1.0)	8 (1.2)	9 (1.3)	10 (0.6)

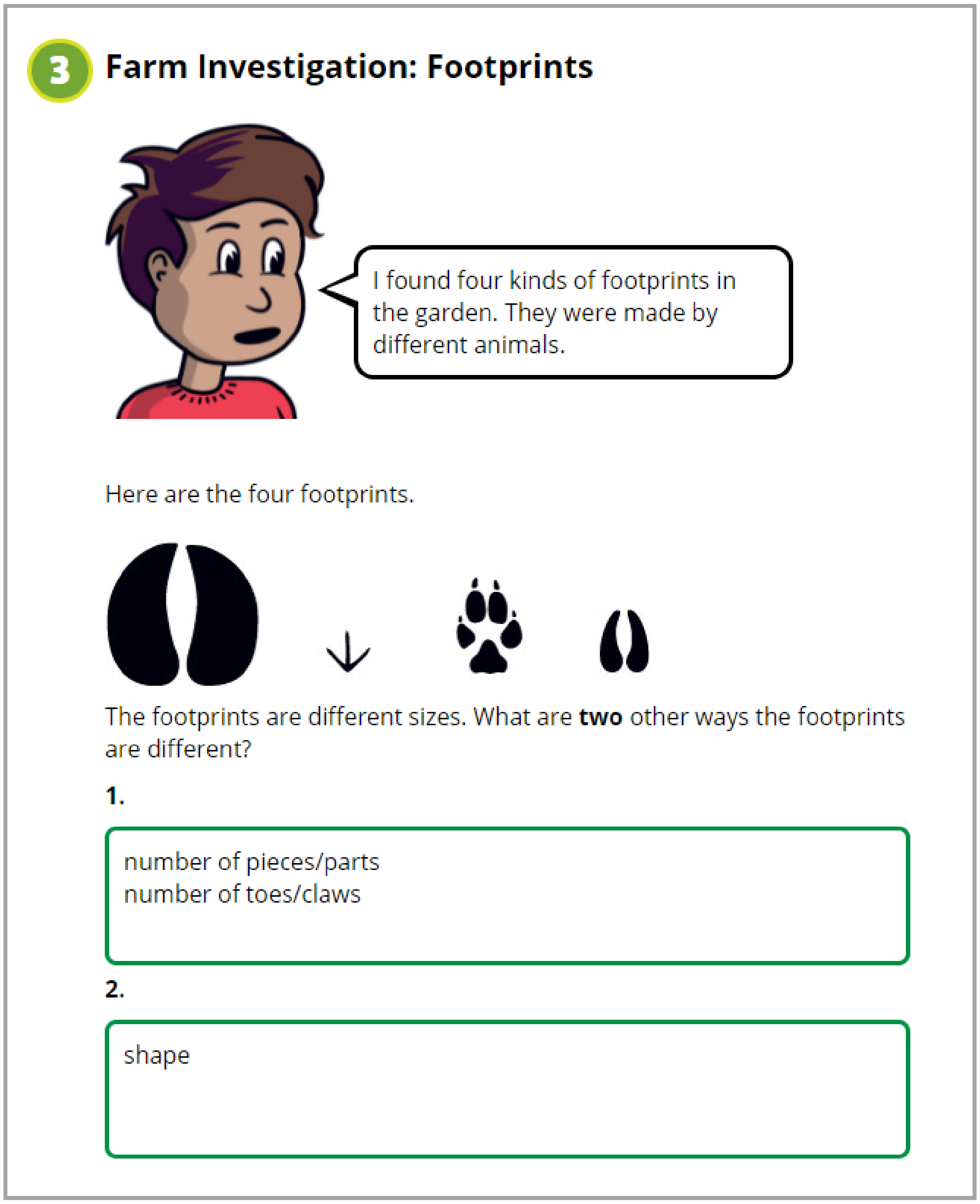
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 3 – Footprints

Regardless of what the students suggested as a clue, George decides to focus on the four different animal footprints that he found in his garden. Students examine the footprints and use their observational skills to provide two ways the footprints differ, other than in size. Responses listing two of the following: number of pieces/parts, number/presence of toes or claws, and shape of footprint were given full credit (1 point).



3 **Farm Investigation: Footprints**

I found four kinds of footprints in the garden. They were made by different animals.

Here are the four footprints.

The footprints are different sizes. What are **two** other ways the footprints are different?

1.

number of pieces/parts
number of toes/claws

2.

shape

Maximum Score Points: 1
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Applying

Exhibit 17 shows the percentage of students in each country that answered this item correctly by describing two differences among the footprints. Average performance on this item (45%) was better than on the previous item, although there was a wide range in performance across countries, from more that 80 percent correct in Korea and Singapore to less than 30 percent in Turkey and France. Girls had higher performance than boys on average internationally (48% vs. 43%).

Farm Investigation Screen 3 – Percent Correct Overall and by Gender

Country	Percent Correct (Gives 2 Correct Differences)		
	Overall Country	Girls	Boys
Korea, Rep. of	82 (2.0)	85 (2.7)	79 (3.0)
Singapore	81 (1.4)	82 (1.9)	80 (1.9)
England	66 (2.1)	63 (3.5)	69 (3.1)
Norway (5)	63 (2.4)	69 (3.3)	57 (3.7)
United States	62 (1.5)	65 (2.3)	58 (2.2)
Sweden	55 (2.7)	59 (3.6)	51 (3.7)
Slovak Republic	55 (2.3)	54 (3.0)	55 (3.2)
Denmark	54 (2.1)	61 (3.2)	45 (3.1)
Hong Kong SAR	51 (3.2)	54 (4.9)	48 (3.3)
Russian Federation	49 (2.2)	48 (3.0)	50 (2.9)
Czech Republic	48 (2.5)	52 (3.3)	44 (3.1)
Canada	48 (1.5)	47 (2.2)	49 (2.2)
Spain	45 (2.1)	47 (3.8)	43 (3.0)
Netherlands	44 (3.0)	44 (3.8)	43 (4.7)
Chinese Taipei	43 (2.3)	45 (3.0)	42 (3.0)
Finland	43 (2.0)	51 (3.0)	35 (3.1)
Lithuania	42 (2.2)	44 (3.2)	40 (3.3)
Malta	42 (2.6)	43 (3.6)	40 (3.3)
Portugal	40 (2.5)	37 (3.2)	42 (3.6)
Georgia	35 (2.7)	38 (4.6)	32 (3.1)
United Arab Emirates	35 (1.0)	38 (1.5)	32 (1.3)
Chile	34 (2.1)	38 (3.1)	32 (2.7)
Qatar	34 (2.6)	39 (3.3)	29 (3.4)
Hungary	34 (2.1)	38 (2.7)	30 (2.7)
Croatia	34 (2.6)	45 (4.6)	23 (2.4)
Italy	33 (2.5)	37 (3.5)	29 (3.3)
Germany	32 (2.5)	37 (3.7)	27 (2.9)
Austria	30 (2.1)	33 (3.1)	28 (3.0)
Turkey (5)	27 (2.2)	25 (3.1)	28 (3.1)
France	22 (2.0)	22 (2.8)	22 (2.9)
International Average	**45 (0.4)**	**48 (0.6)** ▲	**43 (0.6)**
Benchmarking Participants			
Ontario, Canada	58 (2.8)	62 (3.9)	54 (3.9)
Dubai, UAE	53 (1.6)	59 (3.0)	47 (2.1)
Moscow City, Russian Fed.	50 (2.5)	54 (2.7)	45 (3.9)
Madrid, Spain	49 (2.4)	54 (3.5)	43 (3.6)
Abu Dhabi, UAE	24 (1.2)	25 (1.7)	23 (1.8)
Quebec, Canada	23 (2.0)	19 (2.5)	27 (3.3)

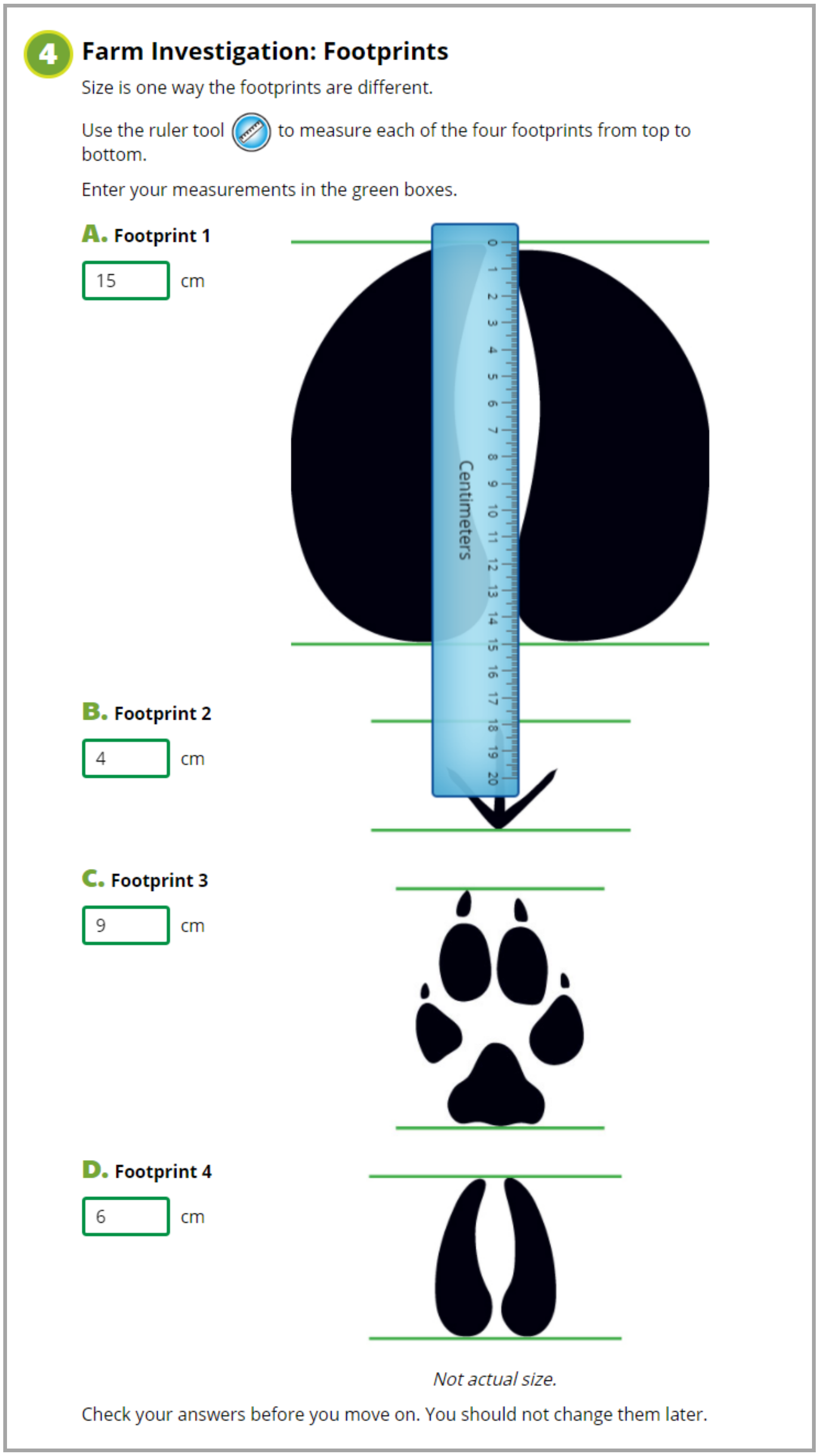
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 4 – Footprints

On Screen 4, students are shown the four footprints and required to demonstrate their ability to make consistent, accurate measurements by using the ruler tool to measure the length in centimeters of each footprint from top to bottom. Students had to correctly position the ruler for each footprint, read the centimeter scale, and record their answers using the number pad. The scoring guide allowed for a tolerance of ± 0.1 cm for student responses. To receive credit (1 point), students had to record correct measurements for all four footprints.



4 **Farm Investigation: Footprints**

Size is one way the footprints are different.

Use the ruler tool to measure each of the four footprints from top to bottom.

Enter your measurements in the green boxes.

A. Footprint 1

15 cm

B. Footprint 2

4 cm

C. Footprint 3

9 cm

D. Footprint 4

6 cm

Not actual size.

Check your answers before you move on. You should not change them later.

Maximum Score Points: 1
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Applying
Science Practice: Generating Evidence

Exhibit 18 presents for each country the percentage of students making all four measurements correctly, overall and separately for boys and girls. This item was well suited to the students' abilities, with 59 percent on average internationally completing the task for all four footprints, and another 11 percent completing three of the four. A smaller percentage of students (11%) gave only three correct measurements for no credit. Performance ranged from 81 percent correct in Chinese Taipei to 29 percent in Qatar. There was no difference overall in the performance of boys and girls. Although this item required the students to manipulate the ruler tool to make the measurements of footprint length, it did not require any science content knowledge, which may have contributed to the reduced gender difference compared to earlier items.

Exhibit 18

Farm Investigation Screen 4 – Percent Correct Overall and by Gender

Country	Percent Correct (Gives 4 Correct Footprint Measurements)			Percent Gives Only 3 Correct Footprint Measurements
	Overall Country	Girls	Boys	
Chinese Taipei	81 (1.9)	79 (2.7)	83 (2.9)	7 (1.2)
Korea, Rep. of	79 (2.1)	81 (2.6)	78 (3.2)	10 (1.3)
Hong Kong SAR	77 (2.1)	76 (3.8)	78 (3.5)	9 (1.6)
Russian Federation	73 (2.1)	74 (2.8)	72 (2.6)	8 (1.2)
Singapore	73 (1.5)	73 (2.3)	72 (2.0)	10 (0.9)
Denmark	72 (2.3)	72 (3.1)	73 (3.6)	11 (1.4)
Czech Republic	70 (2.4)	65 (2.7)	75 (2.9)	9 (1.1)
Slovak Republic	68 (2.0)	66 (2.9)	71 (2.6)	8 (1.1)
Norway (5)	67 (2.5)	67 (3.5)	68 (3.3)	9 (1.4)
England	67 (2.8)	67 (3.5)	67 (3.7)	11 (1.3)
Netherlands	66 (2.6)	63 (3.5)	69 (3.7)	13 (1.6)
Lithuania	66 (2.4)	70 (3.3)	62 (3.5)	14 (1.7)
Sweden	64 (2.8)	68 (4.6)	61 (3.4)	11 (1.7)
Austria	64 (2.2)	63 (3.4)	65 (3.0)	12 (1.4)
Hungary	62 (2.5)	60 (3.4)	64 (3.6)	13 (1.4)
Finland	62 (2.1)	65 (2.7)	59 (3.4)	13 (1.5)
France	60 (2.5)	65 (3.3)	56 (3.7)	11 (1.3)
Germany	59 (2.3)	59 (3.8)	59 (3.0)	14 (1.7)
Portugal	58 (2.1)	55 (3.4)	60 (3.0)	13 (1.7)
Malta	56 (2.4)	53 (3.3)	59 (3.3)	12 (1.5)
Croatia	56 (2.6)	57 (3.2)	54 (3.9)	14 (2.3)
Italy	53 (2.8)	48 (3.6)	56 (4.5)	13 (2.4)
Canada	52 (1.6)	54 (2.4)	51 (2.2)	14 (1.0)
Spain	51 (2.2)	50 (3.7)	52 (2.5)	17 (1.6)
United States	51 (1.9)	47 (2.5)	54 (2.2)	10 (1.0)
Turkey (5)	43 (2.1)	38 (2.9)	49 (3.0)	10 (1.4)
Georgia	35 (3.0)	31 (3.8)	39 (3.7)	10 (1.5)
Chile	33 (2.3)	33 (2.8)	33 (3.1)	16 (1.5)
United Arab Emirates	31 (1.0)	29 (1.3)	33 (1.5)	12 (0.6)
Qatar	29 (1.9)	29 (3.0)	29 (2.5)	8 (1.2)
International Average	**59 (0.4)**	**59 (0.6)**	**60 (0.6)**	**11 (0.3)**
Benchmarking Participants				
Moscow City, Russian Fed.	71 (2.4)	71 (3.1)	71 (3.6)	9 (1.8)
Quebec, Canada	63 (2.6)	65 (3.4)	61 (3.8)	16 (1.9)
Madrid, Spain	56 (2.7)	56 (3.0)	55 (4.5)	16 (1.7)
Ontario, Canada	50 (2.9)	52 (4.4)	48 (3.7)	14 (1.7)
Dubai, UAE	47 (1.9)	43 (2.6)	50 (2.8)	14 (1.3)
Abu Dhabi, UAE	21 (1.2)	19 (1.8)	24 (1.6)	12 (1.1)

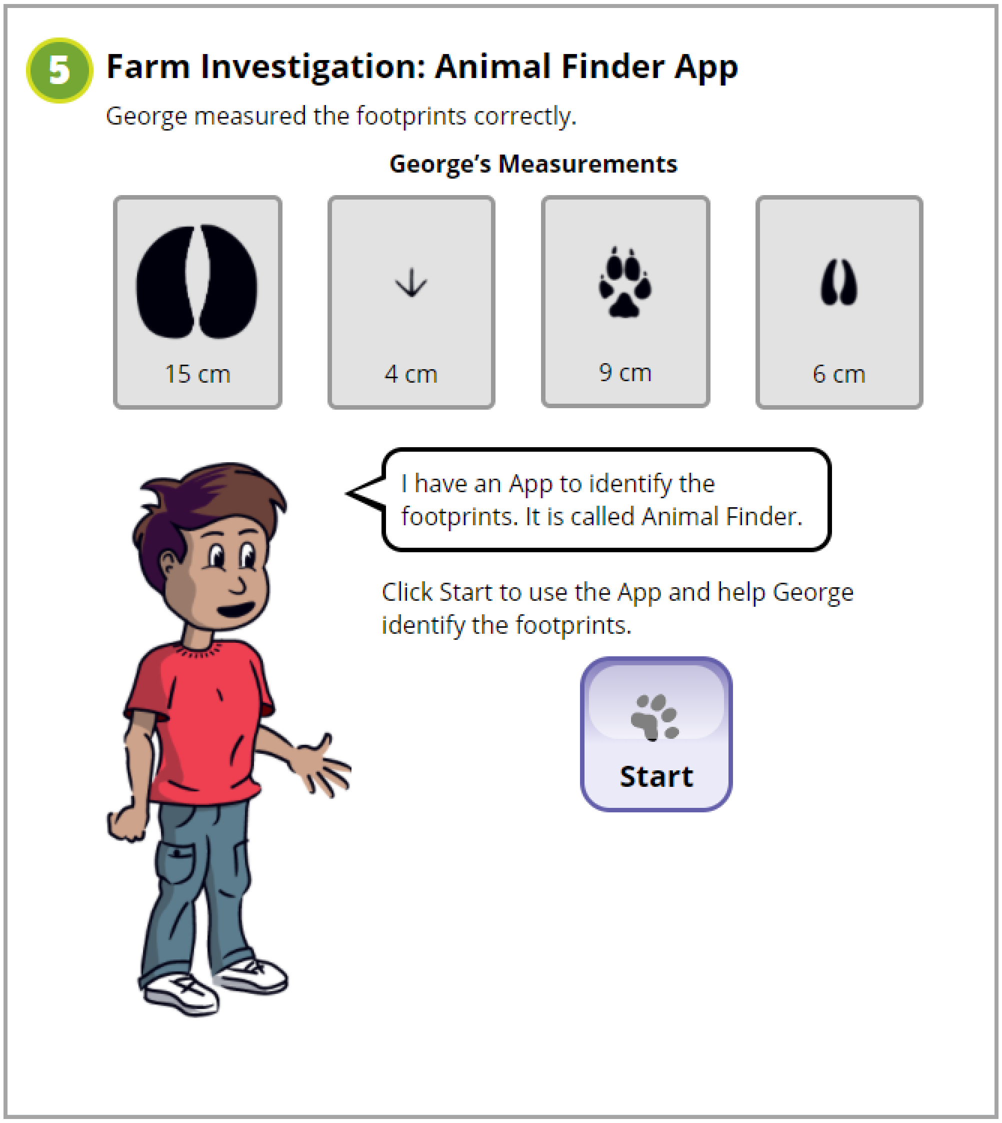
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 5 – Animal Finder App

On Screen 5, students are given George's measurements and told they are correct. George then introduces his "Animal Finder App," which uses a binary decision tree that combines information on the size of each footprint with information on other characteristics of the footprint (number of parts, presence of toes/heel, presence of skin between the toes) to identify the animal that made the footprint. The App can be considered a data-gathering and summarizing tool that students use to combine measurements and observations to reach a conclusion.

Clicking "Start" takes the students to Screen 6 where they use the App to identify the owners of each of the four footprints.



5 Farm Investigation: Animal Finder App

George measured the footprints correctly.

George's Measurements

15 cm	4 cm	9 cm	6 cm

I have an App to identify the footprints. It is called Animal Finder.

Click Start to use the App and help George identify the footprints.

Screen 6 – Animal Finder App

As an example of a correct use of the App, when students click on the first footprint, they are asked:

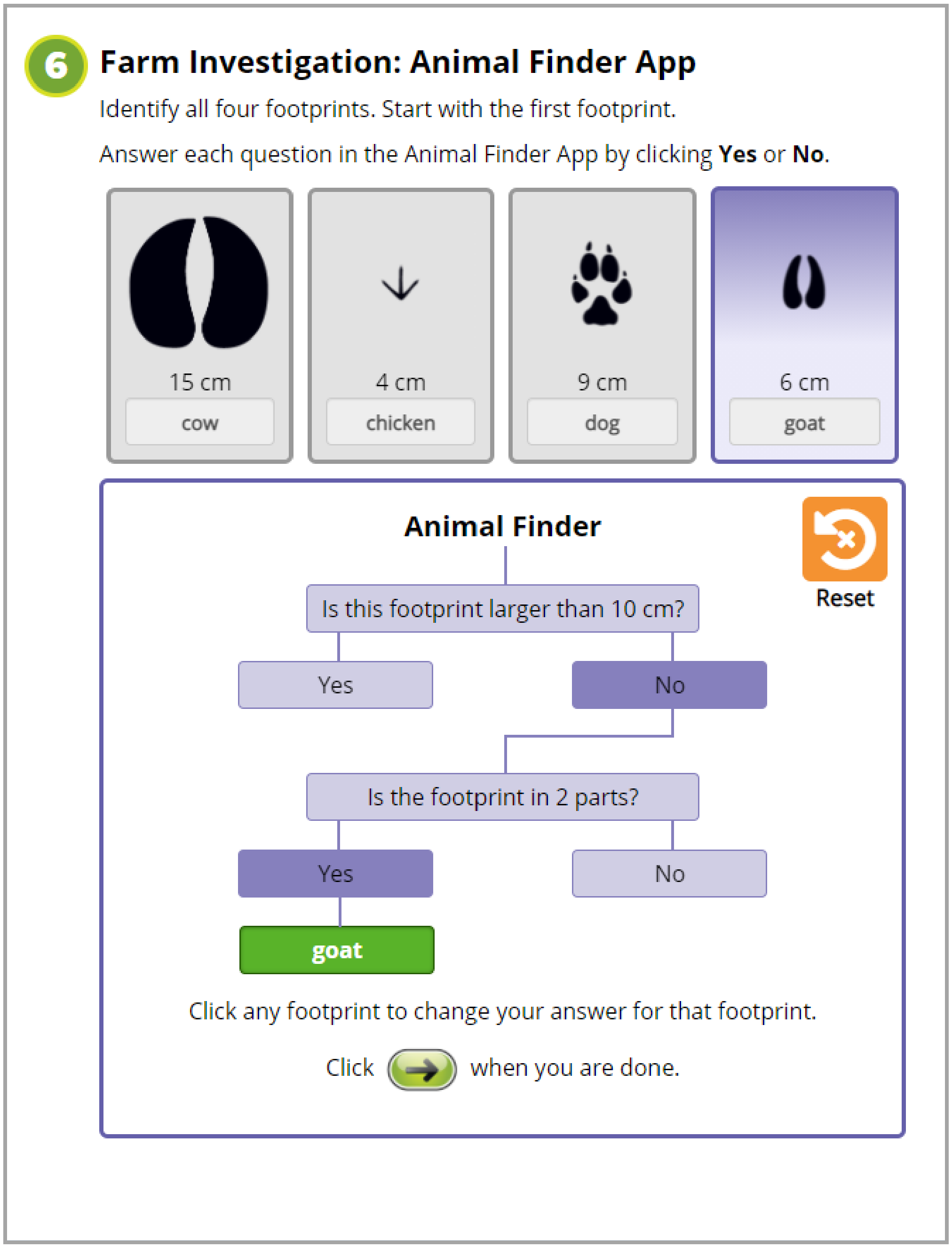
> "Is this footprint larger than 10 cm?" Students click "Yes" or "No".

If they clicked "No", the students are asked:

> "Is the footprint in 2 parts?" Students click "Yes" or "No".

If they clicked "Yes", they are told the footprint belongs to a goat.

Students use the App to identify the animals that made all four footprints. Students who correctly identified all four animals, "cow", "chicken", "dog", and "goat", were given full credit (1 point).



6 Farm Investigation: Animal Finder App

Identify all four footprints. Start with the first footprint.

Answer each question in the Animal Finder App by clicking **Yes** or **No**.

Click any footprint to change your answer for that footprint.

Click ➔ when you are done.

Maximum Score Points: 1
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Applying
Science Practice: Working with Data

Exhibit 19 shows the percentage of students that used the Animal Finder App to identify correctly all four of the animals that made the footprints. More than half the students (54% on average internationally) successfully identified the four animals and a further 19 percent identified three of the four. There was a small gender difference favoring girls (3 percentage points, on average). Similar to the previous item where students had to use the ruler tool to measure the footprints, using the Animal Finder App did not require any science content knowledge of the students, but did require the methodical application of information from observing and measuring the footprints to a series of binary decisions posed by the App.

Exhibit 19

Farm Investigation Screen 6 – Percent Correct Overall and by Gender

Country	Percent Correct (Correctly Identifies 4 Footprints)			Percent Correctly Identifies Only 3 Footprints
	Overall Country	Girls	Boys	
England	72 (2.3)	73 (3.1)	70 (3.0)	14 (1.8)
Korea, Rep. of	67 (2.3)	67 (3.3)	68 (3.0)	19 (1.7)
Norway (5)	67 (2.2)	70 (4.0)	65 (2.8)	12 (1.4)
Singapore	67 (1.8)	70 (2.4)	64 (2.7)	15 (1.3)
Denmark	66 (2.5)	67 (3.5)	65 (3.1)	19 (2.2)
Czech Republic	65 (2.2)	70 (2.9)	59 (3.4)	18 (1.8)
Finland	64 (2.2)	64 (3.2)	64 (2.9)	14 (1.6)
Russian Federation	63 (2.4)	65 (2.7)	60 (3.4)	20 (1.8)
United States	61 (1.7)	61 (2.7)	61 (1.9)	16 (1.3)
Canada	60 (1.8)	63 (2.5)	56 (2.6)	18 (1.2)
Spain	59 (2.8)	60 (4.0)	58 (3.5)	18 (1.5)
Germany	59 (2.4)	62 (3.4)	56 (3.5)	19 (1.9)
Netherlands	59 (2.3)	57 (4.2)	61 (4.3)	18 (1.7)
Slovak Republic	59 (2.6)	60 (3.5)	57 (3.5)	16 (2.1)
Hungary	58 (2.4)	62 (3.7)	54 (3.1)	19 (1.9)
Sweden	55 (2.2)	55 (4.0)	56 (3.7)	17 (2.1)
Austria	54 (2.5)	54 (4.0)	54 (3.3)	16 (1.9)
Hong Kong SAR	53 (2.9)	55 (4.1)	51 (3.6)	24 (2.4)
Lithuania	53 (2.6)	64 (3.1)	42 (3.6)	24 (2.5)
Malta	52 (2.1)	53 (2.8)	51 (3.0)	25 (1.7)
Italy	51 (2.8)	55 (3.8)	47 (4.3)	18 (1.7)
Croatia	47 (2.6)	45 (3.6)	49 (3.7)	30 (2.0)
France	47 (2.6)	49 (3.6)	45 (3.5)	23 (2.2)
Chinese Taipei	47 (2.2)	46 (3.5)	47 (2.7)	24 (1.9)
Portugal	43 (2.6)	42 (3.2)	43 (4.0)	20 (1.6)
Turkey (5)	42 (2.3)	38 (3.1)	46 (3.4)	24 (1.8)
Chile	34 (2.2)	35 (2.8)	33 (3.1)	27 (2.1)
United Arab Emirates	31 (1.0)	31 (1.5)	32 (1.6)	15 (0.7)
Georgia	31 (2.7)	29 (3.6)	32 (3.6)	16 (1.9)
Qatar	27 (2.6)	30 (3.4)	25 (2.9)	15 (2.0)
International Average	**54 (0.4)**	**55 (0.6)** ▲	**52 (0.6)**	**19 (0.3)**
Benchmarking Participants				
Moscow City, Russian Fed.	69 (2.2)	69 (3.1)	68 (3.0)	17 (1.8)
Madrid, Spain	61 (2.9)	60 (4.1)	62 (3.9)	20 (2.3)
Ontario, Canada	59 (2.9)	62 (4.2)	56 (4.0)	19 (2.1)
Quebec, Canada	58 (3.0)	64 (3.7)	50 (4.6)	15 (1.7)
Dubai, UAE	49 (2.0)	49 (2.6)	49 (2.7)	19 (1.7)
Abu Dhabi, UAE	23 (1.2)	21 (1.9)	25 (1.9)	13 (0.9)

▲ Percentage significantly higher than other gender

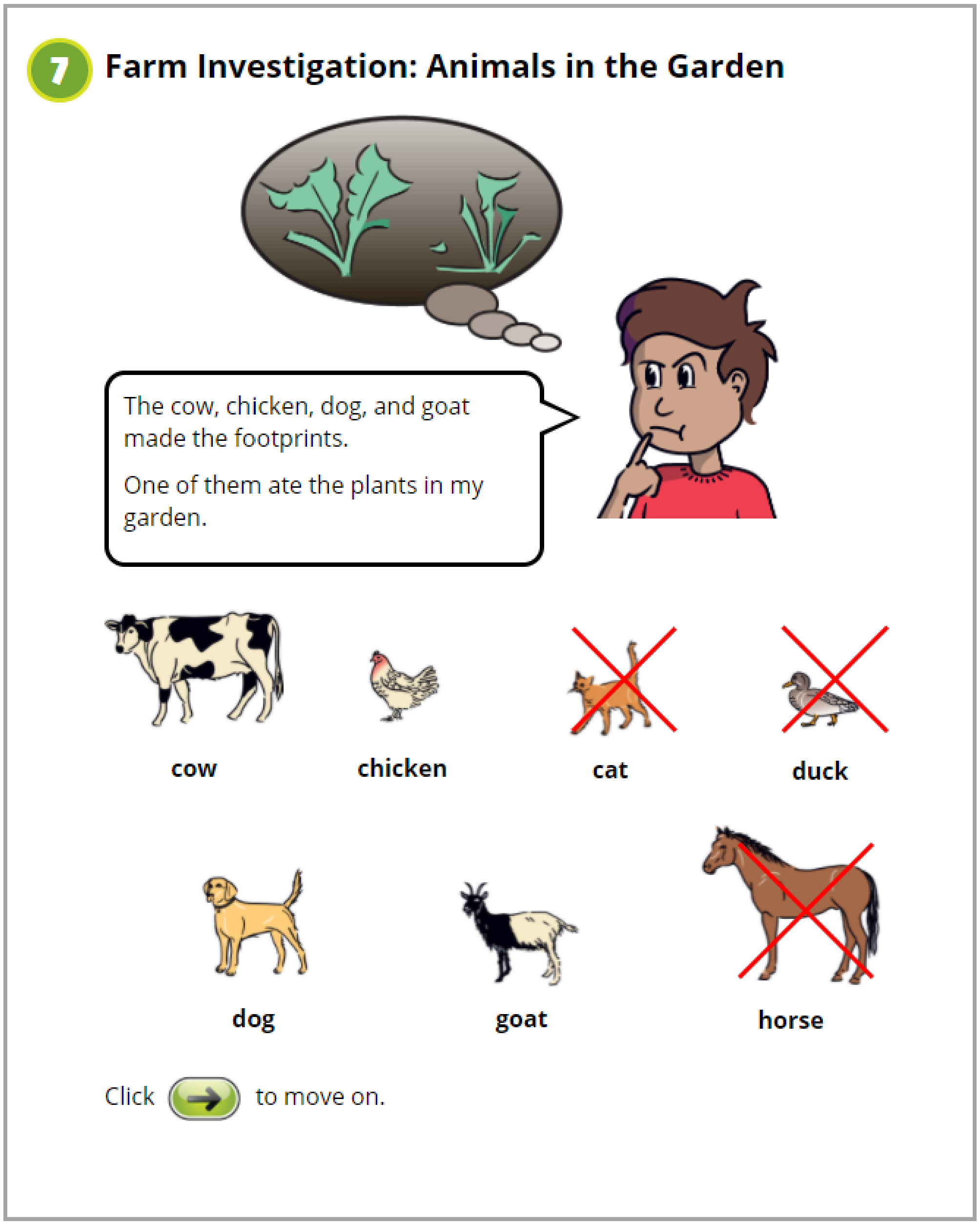
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

IEA
TIMSS & PIRLS
International Study Center
Lynch School of Education
BOSTON COLLEGE

Screen 7 – Animals in the Garden

George confirms on Screen 7 that the footprints were made by the cow, the chicken, the dog, and the goat, and that one of them must have eaten the plants. The cat, duck, and horse are crossed off as possible suspects. By providing the information about which animals made the footprints, this screen sets the next stage of the task and brings all students to the same point of the investigation.



7 **Farm Investigation: Animals in the Garden**

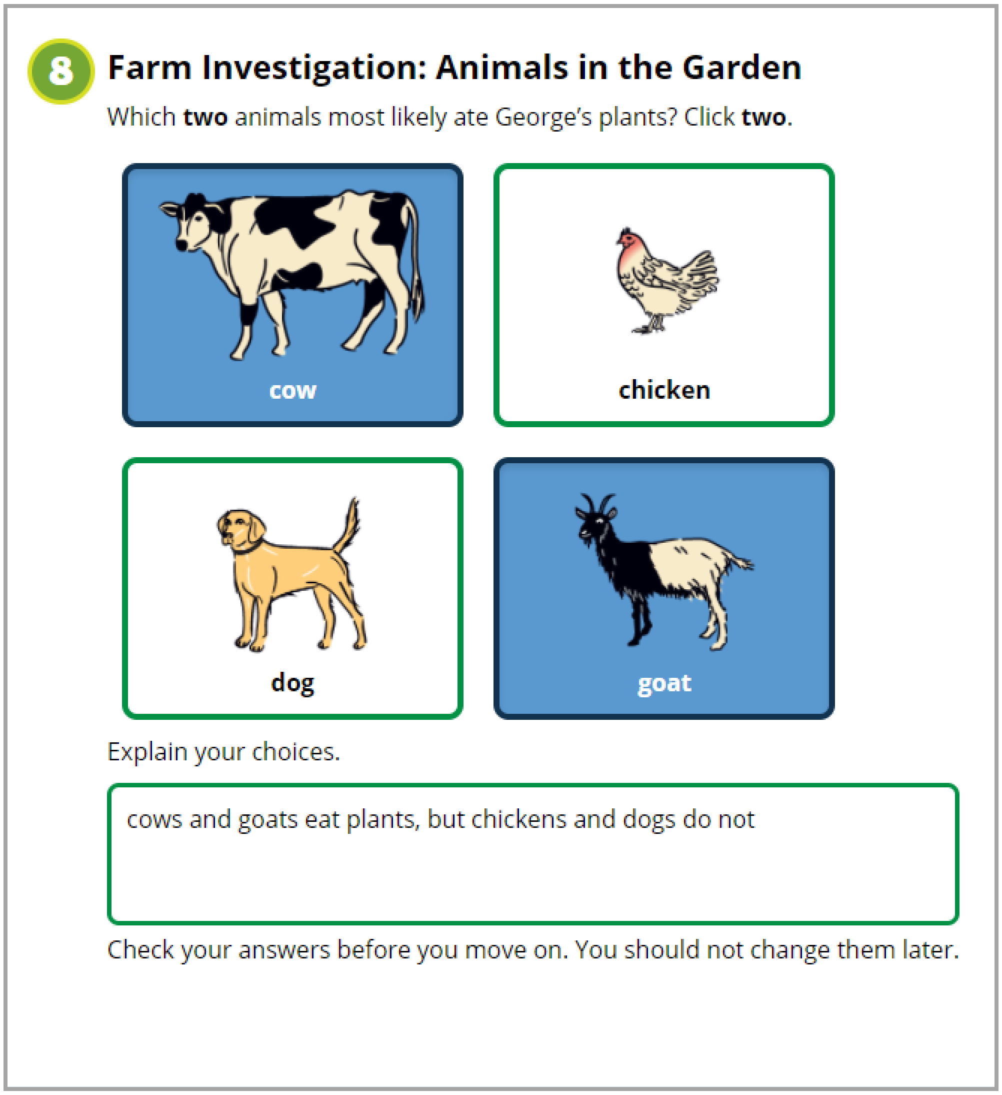
The cow, chicken, dog, and goat made the footprints.

One of them ate the plants in my garden.

Click ➔ to move on.

Screen 8 – Animals in the Garden

On Screen 8, students are asked to identify which two of the four suspect animals (cow, chicken, dog, or goat) were most likely to have eaten the plants and explain their choices. This item assesses basic knowledge of the kinds of foods eaten by domestic animals and does require some life science knowledge. Students were awarded full credit (1 point) for correctly identifying the cow and the goat, and explaining that those animals like to eat plants, or that chickens and dogs prefer to eat other food and do not usually eat plants.



8 **Farm Investigation: Animals in the Garden**

Which **two** animals most likely ate George's plants? Click **two**.

Explain your choices.

cows and goats eat plants, but chickens and dogs do not

Check your answers before you move on. You should not change them later.

Maximum Score Points: 1
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Knowing

Exhibit 20 shows that, on average internationally, more than half the students (52%) correctly identified the cow and goat as likely suspects and provided an acceptable explanation. A further 16 percent identified the two animals but did not provide an explanation, and another 4 percent correctly identified only one animal but without an explanation. There was a small gender difference favoring girls (2 percentage points on average internationally).

Exhibit 20

Farm Investigation Screen 8 – Percent Correct Overall and by Gender

Country	Percent Correct (Selects 2 Correct Animals and Gives Correct Explanation)			Percent Selects 2 Correct Animals but No Correct Explanation	Percent Selects Only 1 Correct Animal and Gives Correct Explanation
	Overall Country	Girls	Boys		
Singapore	70 (1.7)	70 (2.3)	70 (2.2)	8 (1.0)	5 (0.7)
Russian Federation	63 (2.2)	66 (3.2)	61 (2.8)	14 (1.8)	6 (0.8)
Norway (5)	63 (2.4)	64 (3.6)	61 (3.5)	18 (1.9)	2 (0.8)
Korea, Rep. of	60 (2.3)	57 (3.2)	64 (3.0)	7 (1.3)	4 (1.0)
Croatia	60 (2.7)	62 (3.3)	59 (4.0)	7 (1.0)	2 (0.6)
Italy	60 (2.7)	62 (3.9)	57 (3.9)	15 (1.8)	2 (0.7)
Finland	60 (1.9)	62 (2.9)	57 (2.7)	14 (1.3)	3 (0.7)
Lithuania	58 (2.7)	62 (3.6)	54 (3.8)	17 (2.2)	6 (1.3)
Portugal	56 (2.5)	56 (3.8)	56 (3.7)	13 (1.7)	4 (1.0)
Czech Republic	56 (2.5)	58 (3.7)	54 (3.4)	18 (2.1)	6 (1.1)
United States	56 (1.6)	57 (2.4)	55 (2.0)	19 (1.1)	4 (0.5)
Germany	53 (2.6)	58 (3.5)	49 (4.1)	20 (2.3)	3 (0.9)
Slovak Republic	53 (2.3)	55 (3.0)	51 (3.3)	19 (1.6)	4 (0.9)
England	53 (2.9)	53 (3.8)	53 (3.6)	17 (2.0)	2 (0.6)
Netherlands	53 (2.4)	47 (3.3)	58 (3.3)	20 (1.8)	3 (0.8)
Spain	52 (2.3)	51 (3.3)	54 (3.3)	18 (2.0)	5 (1.0)
Sweden	50 (2.6)	55 (4.1)	46 (3.7)	18 (2.3)	3 (1.0)
Denmark	50 (2.9)	53 (4.0)	47 (3.9)	24 (2.5)	1 (0.6)
Canada	50 (1.6)	50 (2.6)	50 (1.8)	20 (1.5)	5 (0.7)
France	48 (2.7)	56 (3.5)	41 (3.7)	21 (2.4)	3 (1.0)
Malta	48 (2.1)	50 (2.9)	46 (3.3)	17 (1.5)	6 (1.0)
Hungary	48 (2.4)	52 (3.2)	43 (2.9)	17 (1.7)	5 (0.9)
Austria	48 (2.5)	47 (3.6)	49 (3.5)	23 (2.0)	5 (0.8)
Turkey (5)	46 (2.8)	41 (3.1)	51 (3.9)	21 (1.9)	3 (0.8)
Hong Kong SAR	46 (2.8)	47 (3.6)	45 (4.5)	12 (1.8)	5 (0.9)
Chile	45 (2.0)	42 (3.2)	47 (2.7)	10 (1.4)	4 (0.8)
United Arab Emirates	41 (1.1)	42 (1.5)	39 (1.6)	9 (0.5)	6 (0.6)
Chinese Taipei	40 (1.8)	42 (3.0)	38 (3.5)	7 (1.3)	3 (0.8)
Qatar	38 (2.5)	42 (3.5)	34 (3.2)	13 (1.5)	7 (0.9)
Georgia	36 (2.9)	38 (3.6)	34 (3.5)	22 (2.5)	6 (1.2)
International Average	**52 (0.4)**	**53 (0.6)** ▲	**51 (0.6)**	**16 (0.3)**	**4 (0.2)**
Benchmarking Participants					
Moscow City, Russian Fed.	69 (2.1)	74 (3.1)	63 (3.3)	14 (1.9)	6 (1.1)
Madrid, Spain	58 (2.3)	58 (3.6)	58 (3.8)	17 (2.1)	5 (1.3)
Dubai, UAE	57 (2.6)	60 (3.9)	55 (2.5)	7 (0.8)	7 (1.1)
Quebec, Canada	56 (2.9)	56 (4.1)	57 (3.5)	16 (2.0)	4 (1.0)
Ontario, Canada	49 (2.2)	50 (3.9)	49 (3.2)	20 (2.5)	5 (1.2)
Abu Dhabi, UAE	32 (1.4)	31 (2.0)	33 (2.1)	5 (0.8)	4 (0.8)

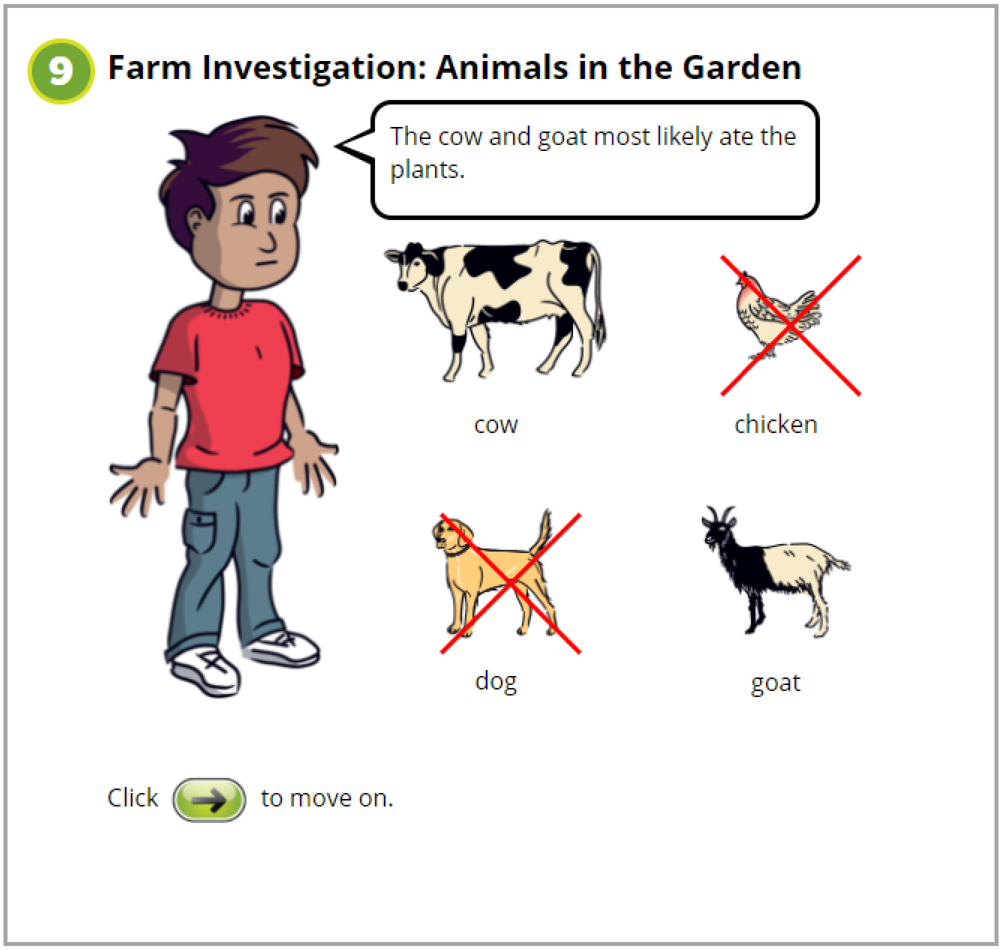
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 9 – Animals in the Garden

George decides on Screen 9 that the cow or the goat most likely ate the plants, and crosses off the dog and the chicken as possible suspects.



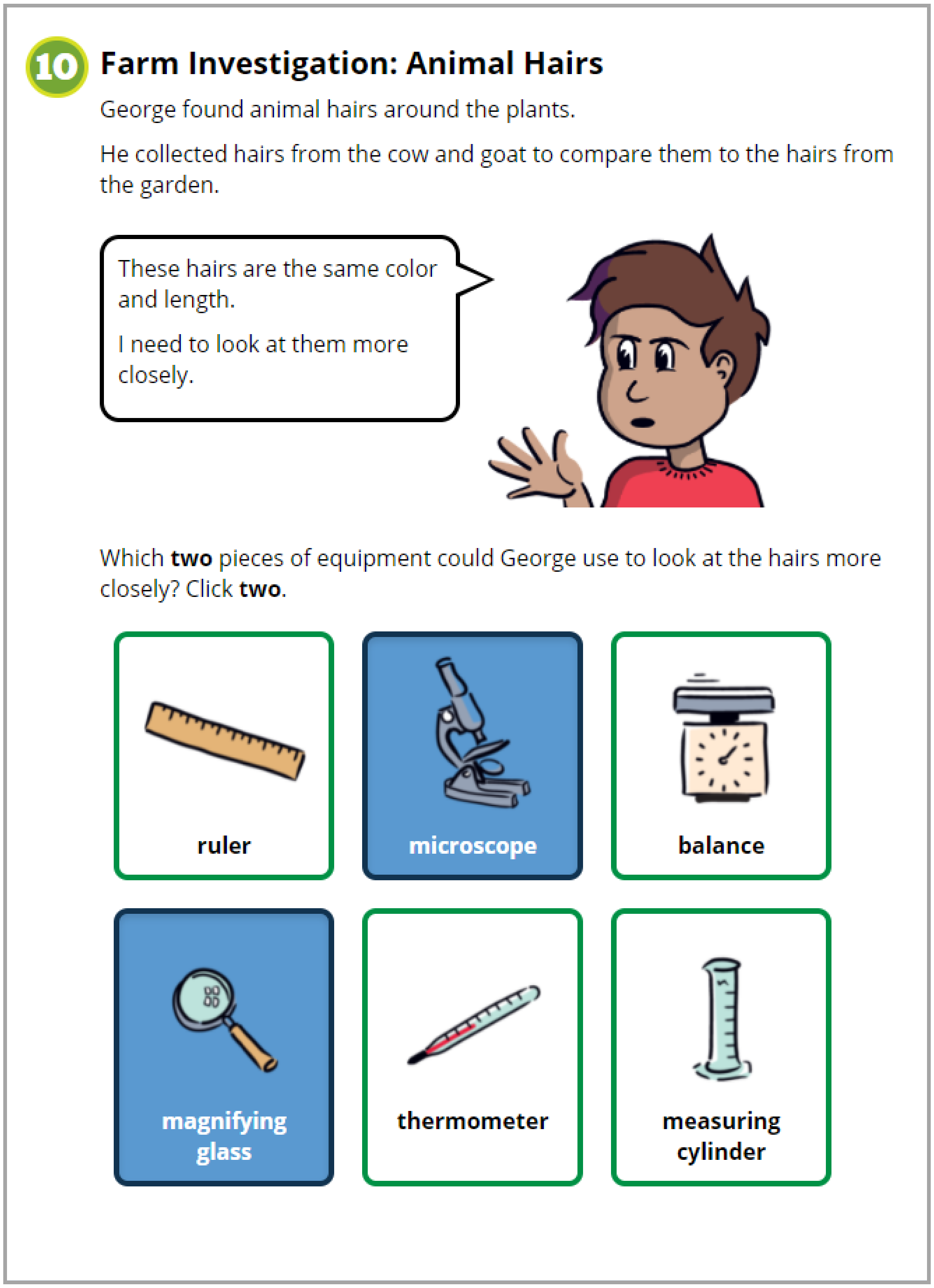
9 Farm Investigation: Animals in the Garden

The cow and goat most likely ate the plants.

Screen 10 – Animal Hairs

Continuing his investigation, George discovers animal hairs on the ground near his plants. He decides to collect hairs from the cow and the goat to compare them to the hairs from the garden. He notes that the hairs are the same color and length, so he needs to examine them more closely.

Students are shown six pieces of measuring equipment (ruler, microscope, balance, magnifying glass, thermometer, and measuring cylinder) and asked to click on the two that could be used to look at the hairs more closely. This item assesses familiarity with common measuring instruments and basic knowledge of the purposes for which they are intended. Students correctly identifying the microscope and magnifying glass were awarded full credit (1 point).



10 **Farm Investigation: Animal Hairs**

George found animal hairs around the plants.

He collected hairs from the cow and goat to compare them to the hairs from the garden.

These hairs are the same color and length.

I need to look at them more closely.

Which **two** pieces of equipment could George use to look at the hairs more closely? Click **two**.

Maximum Score Points: 1
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Knowing

Exhibit 21 shows the percentage of students in each country that correctly selected the microscope and magnifying glass. In general, students performed well on this item, with 64 percent correctly identifying the two instruments, and a further 12 percent identifying one instrument only. Girls performed better than boys on average internationally (by 4 percentage points).

Exhibit 21

Farm Investigation Screen 10 – Percent Correct Overall and by Gender

Country	Percent Correct (Selects 2 Correct Instruments)			Percent Selects Only 1 Correct Instrument
	Overall Country	Girls	Boys	
Slovak Republic	76 (1.9)	81 (2.5)	72 (3.0)	9 (1.3)
Portugal	74 (1.9)	78 (3.1)	71 (2.6)	10 (1.5)
Italy	73 (2.7)	72 (3.3)	74 (3.7)	13 (2.3)
Netherlands	73 (2.1)	75 (3.2)	71 (2.7)	11 (1.5)
Lithuania	72 (2.2)	77 (3.1)	68 (2.9)	13 (1.7)
Denmark	71 (2.5)	74 (3.5)	68 (3.5)	9 (1.5)
Germany	71 (2.1)	73 (3.2)	69 (3.4)	9 (1.6)
Croatia	71 (2.8)	76 (2.4)	67 (4.7)	4 (1.1)
Czech Republic	71 (2.2)	71 (3.4)	70 (3.1)	11 (1.7)
Austria	69 (2.1)	69 (3.0)	70 (3.1)	10 (1.4)
England	69 (2.5)	68 (3.4)	69 (3.4)	11 (1.4)
Spain	68 (2.5)	68 (3.8)	67 (3.0)	14 (1.4)
Russian Federation	67 (2.3)	65 (2.9)	69 (3.2)	15 (1.8)
Malta	66 (2.2)	68 (2.6)	65 (3.2)	9 (1.3)
Norway (5)	65 (2.0)	68 (3.7)	63 (2.9)	11 (1.6)
Singapore	64 (1.7)	67 (2.4)	62 (2.3)	9 (1.1)
Hungary	64 (2.3)	70 (3.4)	57 (3.1)	12 (1.8)
Hong Kong SAR	63 (2.6)	63 (3.0)	63 (4.1)	12 (1.6)
Korea, Rep. of	63 (2.4)	65 (3.6)	61 (3.4)	8 (1.3)
United States	63 (1.6)	64 (2.3)	61 (2.2)	15 (1.1)
Chinese Taipei	62 (2.2)	64 (3.3)	60 (3.2)	7 (1.2)
Finland	62 (2.1)	58 (3.2)	65 (2.7)	7 (1.0)
Turkey (5)	61 (2.2)	61 (3.5)	62 (3.1)	16 (1.5)
Chile	59 (2.6)	65 (3.7)	54 (3.1)	20 (2.1)
France	58 (2.6)	63 (3.3)	54 (3.7)	14 (1.9)
Canada	58 (1.9)	65 (2.5)	51 (2.6)	10 (1.0)
Sweden	49 (3.1)	49 (4.2)	50 (3.9)	9 (1.5)
Georgia	48 (2.8)	55 (4.5)	42 (3.6)	29 (2.7)
United Arab Emirates	47 (1.3)	51 (1.9)	44 (1.9)	22 (1.0)
Qatar	42 (2.2)	46 (3.0)	39 (2.7)	20 (1.9)
International Average	**64 (0.4)**	**66 (0.6)** ▲	**62 (0.6)**	**12 (0.3)**
Benchmarking Participants				
Moscow City, Russian Fed.	72 (1.9)	75 (2.9)	69 (2.6)	11 (1.5)
Madrid, Spain	69 (2.4)	70 (3.0)	67 (3.5)	11 (1.6)
Dubai, UAE	60 (2.4)	65 (3.3)	55 (2.8)	17 (1.6)
Ontario, Canada	57 (3.0)	66 (4.2)	49 (3.8)	11 (1.8)
Quebec, Canada	55 (3.1)	61 (4.0)	48 (4.1)	7 (1.1)
Abu Dhabi, UAE	40 (1.5)	41 (2.7)	38 (2.6)	21 (1.5)

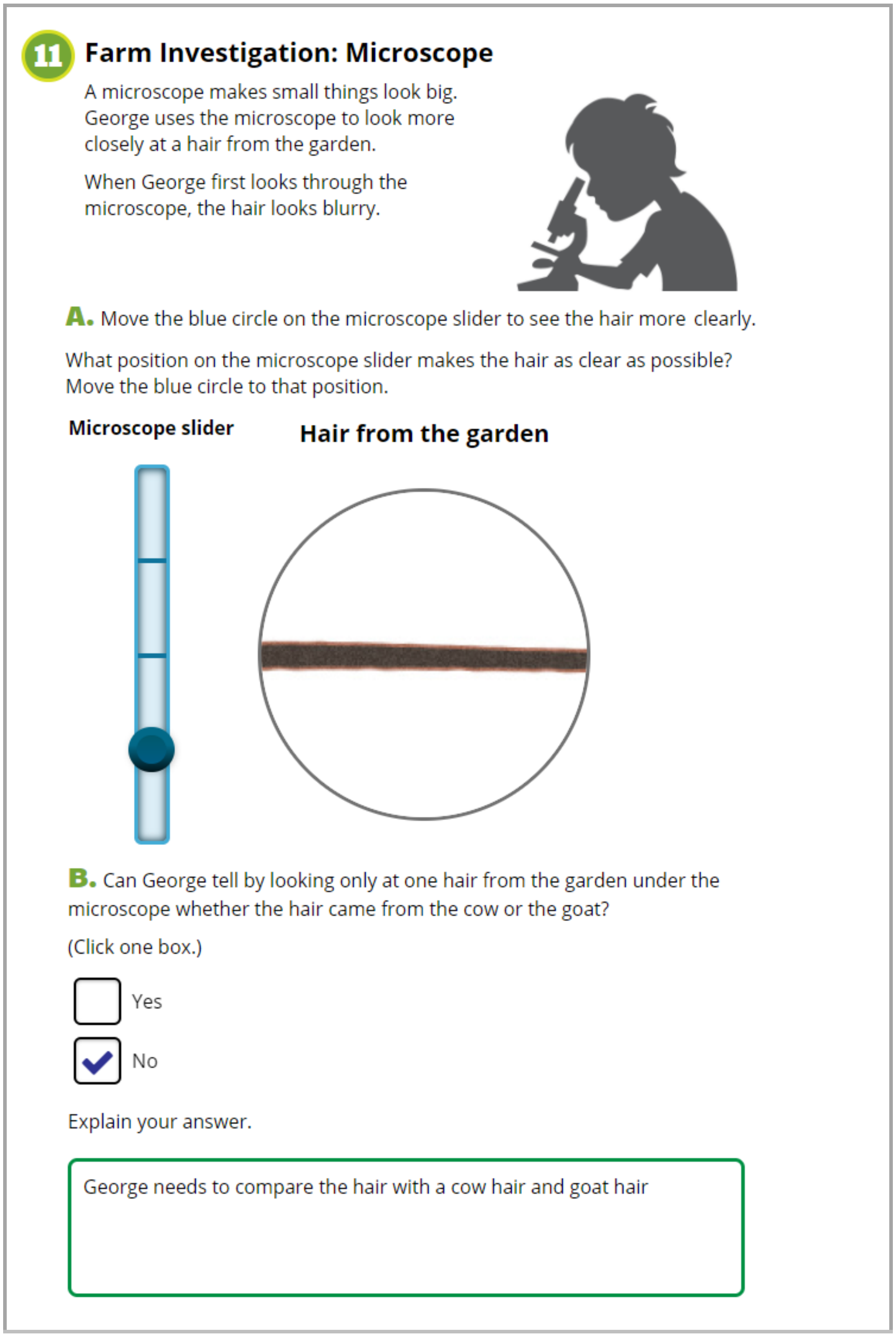
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 11 – Microscope

Students are told that a microscope can make small things look big, and so George uses a microscope to look more closely at a hair from the garden. When George first looks through the microscope, the hair looks blurry.



11 Farm Investigation: Microscope

A microscope makes small things look big. George uses the microscope to look more closely at a hair from the garden.

When George first looks through the microscope, the hair looks blurry.

A. Move the blue circle on the microscope slider to see the hair more clearly.

What position on the microscope slider makes the hair as clear as possible? Move the blue circle to that position.

Microscope slider

Hair from the garden

B. Can George tell by looking only at one hair from the garden under the microscope whether the hair came from the cow or the goat?

(Click one box.)

□ Yes

☑ No

Explain your answer.

George needs to compare the hair with a cow hair and goat hair

	Item 11A	Item 11B
Maximum Score Points:	1	1
Content Domain:	Life Science	Life Science
Topic Area:	Characteristics and Life Processes of Organisms	Characteristics and Life Processes of Organisms
Cognitive Domain:	Knowing	Knowing
Science Practice:	Generating Evidence	Generating Evidence

In 11A, students are told that they can move the blue circle on the microscope slider to see the hair more clearly, and are asked to find the position on the slider that makes the hair as clear as possible. This simulates the procedure for focusing a microscope in the laboratory, and assesses students' ability to select the correct setting. There are five possible positions on the slider, only one of which shows the hair clearly. Students choosing the correct position were awarded full credit (1 point).

Exhibit 22 shows the percentages of students in each country that were able to focus the microscope. This was a very straightforward task, intended to introduce the students to the operation of the microscope, and was completed correctly by 61 percent of the students, on average internationally. Girls performed a little better than boys on this item, on average (63% vs. 59% correct).

Farm Investigation Screen 11A – Percent Correct Overall and by Gender

Country	Percent Correct (Focuses Microscope Correctly)		
	Overall Country	Girls	Boys
Finland	75 (1.9)	75 (2.7)	75 (2.7)
Denmark	73 (2.0)	75 (2.6)	71 (2.8)
Czech Republic	73 (2.0)	73 (2.9)	72 (3.2)
Russian Federation	73 (2.0)	72 (3.2)	73 (2.8)
England	72 (2.3)	72 (3.2)	71 (3.6)
Lithuania	72 (2.3)	71 (3.4)	72 (3.0)
Germany	69 (2.2)	73 (3.2)	65 (3.2)
United States	68 (1.4)	72 (2.3)	65 (1.8)
Austria	68 (2.3)	69 (3.1)	67 (3.6)
Croatia	66 (2.9)	73 (4.1)	61 (3.9)
Netherlands	66 (2.6)	66 (3.5)	66 (3.8)
Hungary	65 (2.3)	65 (3.0)	64 (3.2)
Slovak Republic	64 (2.2)	66 (3.0)	63 (3.6)
Turkey (5)	63 (2.1)	58 (3.4)	68 (3.0)
Sweden	62 (2.6)	65 (3.9)	59 (3.4)
Norway (5)	62 (2.6)	68 (3.6)	56 (3.5)
France	62 (2.0)	66 (3.3)	58 (3.2)
Malta	58 (2.3)	64 (3.0)	53 (3.3)
Canada	58 (1.6)	62 (2.4)	54 (2.1)
Italy	56 (2.6)	56 (4.1)	56 (3.6)
Singapore	56 (1.7)	59 (2.4)	53 (2.5)
Chile	55 (2.3)	55 (3.9)	55 (3.4)
Spain	54 (2.6)	55 (3.2)	53 (3.3)
Portugal	52 (2.8)	56 (3.4)	48 (3.8)
Georgia	49 (2.8)	52 (3.9)	47 (3.6)
Korea, Rep. of	49 (2.3)	52 (2.9)	45 (2.9)
United Arab Emirates	46 (1.1)	47 (1.7)	45 (1.6)
Qatar	45 (2.6)	50 (3.4)	40 (3.7)
Chinese Taipei	38 (2.1)	38 (3.2)	39 (3.0)
Hong Kong SAR	- -	- -	- -
International Average	**61 (0.4)**	**63 (0.6)** ▲	**59 (0.6)**
Benchmarking Participants			
Moscow City, Russian Fed.	73 (2.0)	75 (2.7)	70 (3.1)
Quebec, Canada	66 (2.8)	69 (3.5)	63 (4.2)
Dubai, UAE	57 (2.0)	58 (2.9)	56 (2.6)
Madrid, Spain	55 (2.6)	60 (3.4)	50 (3.5)
Ontario, Canada	53 (2.7)	59 (4.1)	48 (3.3)
Abu Dhabi, UAE	44 (1.8)	44 (2.9)	44 (2.6)

▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.
A dash (–) indicates comparable data not available.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

To probe their deductive reasoning skills, in 11B students are asked if George can tell by looking only at one hair from the garden under the microscope whether the hair came from the cow or the goat. To be awarded credit (1 point), students had to indicate that George could not tell from one hair only, and describe the additional information needed or the method by which the additional information could be obtained. For example, "No—George needs to compare the hair he found to hairs from each of the animals and see which match."

Exhibit 23 shows that students found this item to be very difficult, with only 13 percent of students on average internationally providing an acceptable explanation.

Farm Investigation Screen 11B – Percent Correct Overall and by Gender

Country	Percent Correct (Correct Explanation Describing Need for Additional Information)		
	Overall Country	Girls	Boys
Denmark	31 (2.4)	39 (3.6)	24 (3.1)
Russian Federation	23 (2.1)	25 (3.1)	22 (2.4)
Korea, Rep. of	23 (1.8)	21 (2.8)	24 (2.4)
Norway (5)	22 (2.4)	27 (3.7)	17 (2.8)
Croatia	22 (2.3)	31 (4.3)	13 (2.0)
Turkey (5)	21 (1.7)	26 (2.6)	17 (2.3)
Netherlands	21 (2.2)	26 (3.9)	16 (3.1)
England	19 (2.1)	22 (3.4)	17 (2.7)
Singapore	17 (1.3)	19 (2.0)	14 (1.8)
Canada	15 (1.3)	17 (1.9)	13 (2.0)
Lithuania	14 (1.9)	15 (2.6)	14 (2.3)
Hungary	14 (1.8)	17 (2.7)	10 (2.2)
Malta	13 (1.5)	16 (2.4)	9 (1.6)
Spain	12 (1.6)	12 (2.0)	13 (2.4)
United States	12 (1.1)	14 (1.6)	11 (1.5)
Finland	12 (1.5)	14 (2.4)	10 (2.1)
Slovak Republic	11 (1.4)	13 (2.0)	10 (2.0)
Sweden	11 (2.0)	10 (2.1)	13 (3.2)
Chinese Taipei	9 (1.4)	12 (2.0)	7 (2.0)
Germany	9 (1.6)	12 (2.4)	6 (1.6)
Austria	9 (1.1)	11 (2.1)	7 (1.6)
Czech Republic	9 (1.2)	10 (1.7)	8 (1.8)
Hong Kong SAR	8 (1.3)	7 (1.3)	8 (2.4)
France	7 (1.6)	8 (2.8)	5 (1.4)
United Arab Emirates	5 (0.5)	5 (0.7)	5 (0.5)
Portugal	5 (1.2)	6 (2.3)	4 (1.2)
Georgia	4 (1.1)	3 (1.4)	4 (1.6)
Qatar	4 (1.0)	5 (1.6)	2 (1.2)
Chile	3 (1.0)	5 (1.6)	2 (1.0)
Italy	3 (0.9)	3 (1.2)	3 (1.3)
International Average	**13 (0.3)**	**15 (0.5)** ▲	**11 (0.4)**
Benchmarking Participants			
Ontario, Canada	19 (2.7)	24 (3.7)	15 (3.8)
Moscow City, Russian Fed.	15 (1.5)	16 (2.3)	13 (2.5)
Dubai, UAE	11 (1.2)	10 (1.8)	12 (1.5)
Quebec, Canada	9 (1.3)	9 (2.0)	9 (1.8)
Madrid, Spain	8 (1.3)	10 (1.8)	6 (1.8)
Abu Dhabi, UAE	2 (0.5)	2 (0.7)	3 (0.7)

▲ Percentage significantly higher than other gender

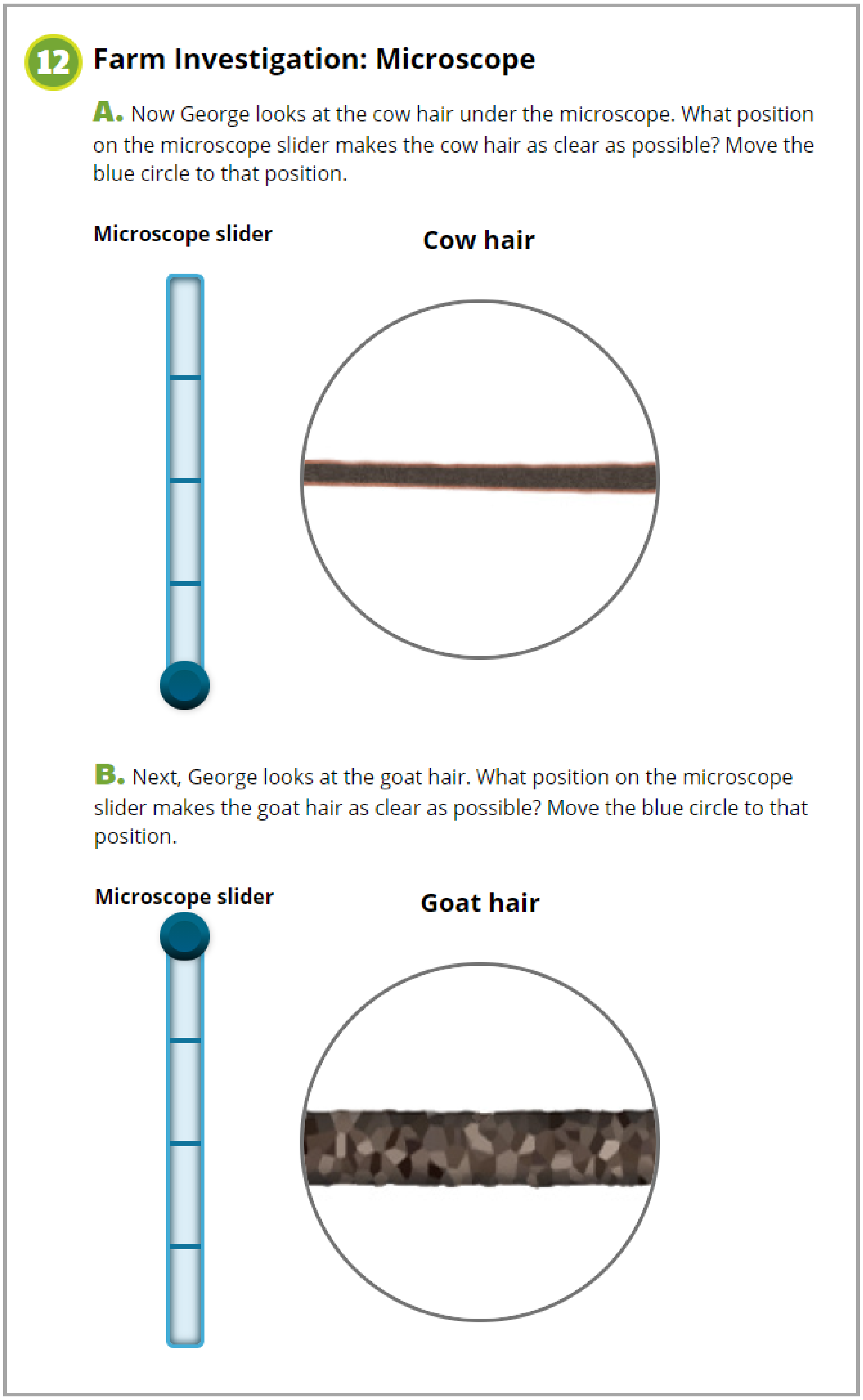
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 12 – Microscope

On Screen 12, George uses the microscope to examine a cow hair and a goat hair. Students are asked to move the slider to make the images as clear as possible.

In 12A, students adjust the slider to make the cow hair as clear as possible, and in 12B to make the goat hair as clear as possible. Students were awarded credit (1 point) for correctly positioning the slider for both animal hairs.



12 Farm Investigation: Microscope

A. Now George looks at the cow hair under the microscope. What position on the microscope slider makes the cow hair as clear as possible? Move the blue circle to that position.

B. Next, George looks at the goat hair. What position on the microscope slider makes the goat hair as clear as possible? Move the blue circle to that position.

Maximum Score Points: 1
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Knowing
Science Practice: Generating Evidence

Positioning the microscope slider is a relatively straightforward task, which students performed reasonably well, shown in Exhibit 24. On average internationally, two-thirds of the students (66%) focused the microscope correctly on both animal hairs. In general, focusing the microscope was something the students were able to do for both hairs or not at all. Only 4 percent on average were able to focus on the cow hair but not the goat hair, and just 8 percent on the goat hair but not the cow hair. There was little difference between girls and boys in performance on this item.

Exhibit 24

Farm Investigation Screen 12 – Percent Correct Overall and by Gender

Country	Percent Correct (Focuses Microscope Correctly for 2 Hairs)			Percent Focuses Microscope Correctly for Cow Hair Only	Percent Focuses Microscope Correctly for Goat Hair Only
	Overall Country	Girls	Boys		
Denmark	83 (2.0)	84 (2.8)	83 (3.1)	3 (0.9)	4 (1.0)
Finland	81 (1.7)	81 (2.2)	81 (2.7)	3 (0.7)	6 (1.0)
Netherlands	78 (2.0)	80 (2.7)	76 (3.1)	2 (0.8)	9 (1.5)
England	76 (2.1)	79 (2.8)	74 (3.0)	3 (1.1)	7 (1.4)
Norway (5)	75 (2.6)	77 (3.5)	73 (3.7)	2 (0.7)	8 (1.6)
Malta	75 (2.2)	75 (2.8)	74 (2.9)	3 (0.7)	7 (1.4)
Sweden	74 (2.4)	79 (2.8)	70 (3.4)	3 (0.8)	7 (1.2)
Croatia	74 (2.5)	74 (4.2)	74 (2.9)	3 (0.8)	6 (1.4)
Czech Republic	74 (2.4)	76 (2.4)	72 (3.6)	3 (0.6)	9 (1.4)
United States	74 (1.5)	73 (2.2)	74 (2.1)	3 (0.7)	8 (0.9)
Singapore	72 (1.6)	73 (2.3)	70 (2.1)	5 (0.9)	8 (1.0)
France	70 (2.3)	72 (3.1)	68 (3.2)	1 (0.5)	7 (1.3)
Austria	70 (2.0)	69 (3.0)	71 (2.8)	7 (1.1)	6 (1.0)
Lithuania	70 (2.6)	70 (3.6)	70 (3.6)	4 (0.8)	10 (1.6)
Hungary	69 (1.9)	66 (2.9)	72 (2.7)	3 (0.5)	10 (1.3)
Russian Federation	69 (2.0)	68 (3.0)	71 (3.1)	2 (0.6)	11 (1.4)
Slovak Republic	69 (2.3)	67 (3.5)	70 (3.8)	3 (0.8)	8 (1.4)
Spain	68 (2.3)	64 (3.0)	71 (2.6)	4 (0.9)	11 (1.7)
Turkey (5)	66 (2.6)	62 (3.6)	69 (3.2)	3 (0.9)	8 (1.5)
Germany	65 (2.4)	68 (3.8)	62 (3.8)	4 (1.1)	8 (1.3)
Korea, Rep. of	64 (2.2)	65 (3.4)	62 (3.0)	3 (0.6)	10 (1.9)
Canada	62 (1.6)	63 (2.6)	61 (2.4)	7 (1.1)	8 (0.9)
Chinese Taipei	62 (2.2)	63 (3.3)	60 (3.0)	5 (1.0)	7 (1.1)
Hong Kong SAR	57 (2.9)	53 (4.3)	61 (3.4)	3 (1.0)	8 (1.4)
Portugal	56 (2.6)	60 (3.9)	54 (3.3)	6 (1.1)	6 (1.2)
Chile	55 (2.3)	57 (3.8)	53 (3.5)	5 (1.2)	10 (1.6)
Italy	54 (2.9)	59 (4.0)	49 (3.7)	9 (1.6)	7 (1.3)
United Arab Emirates	49 (1.0)	46 (1.6)	52 (1.4)	6 (0.5)	8 (0.5)
Qatar	43 (2.9)	44 (3.9)	43 (3.9)	7 (1.5)	11 (1.6)
Georgia	39 (2.5)	35 (4.0)	42 (3.6)	5 (1.3)	10 (1.5)
International Average	**66 (0.4)**	**67 (0.6)**	**66 (0.6)**	**4 (0.2)**	**8 (0.2)**
Benchmarking Participants					
Madrid, Spain	75 (2.1)	76 (3.0)	73 (3.1)	3 (0.8)	9 (1.5)
Moscow City, Russian Fed.	73 (1.6)	73 (2.4)	74 (2.9)	2 (0.5)	9 (1.1)
Quebec, Canada	67 (3.0)	66 (4.4)	69 (4.0)	6 (1.4)	8 (1.4)
Dubai, UAE	65 (1.6)	62 (2.5)	67 (2.0)	5 (1.1)	8 (0.9)
Ontario, Canada	59 (2.9)	60 (4.7)	58 (4.0)	8 (1.8)	6 (1.7)
Abu Dhabi, UAE	45 (1.8)	41 (2.3)	49 (2.3)	5 (0.9)	8 (0.9)

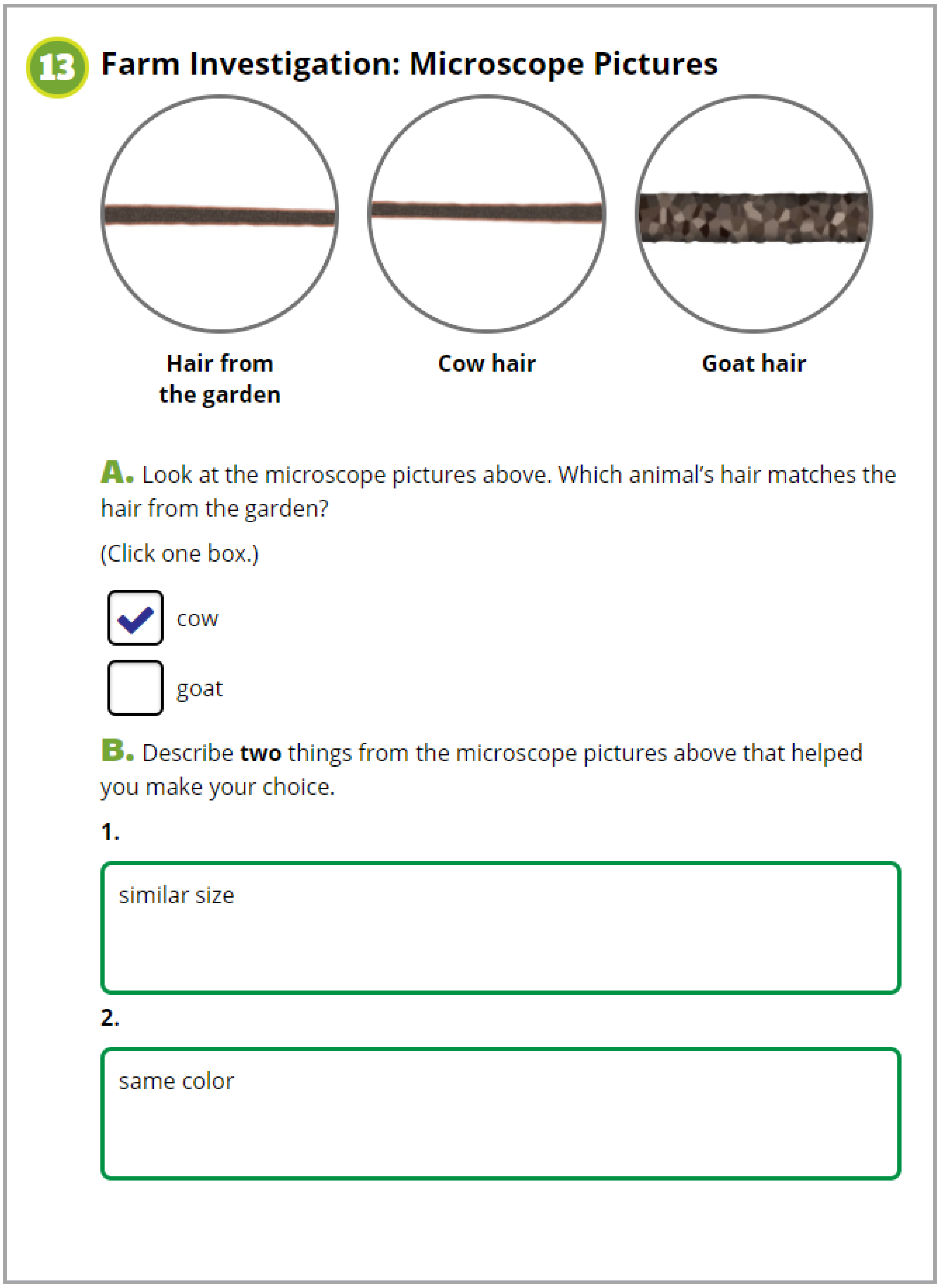
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 13 – Microscope Pictures

On this screen, students are shown microscope pictures of the hair from the garden and of the cow and goat hairs, and asked to identify the animal whose hair matched the hair from the garden, describing two things from the pictures that helped them make their choice. This item requires students to use their observational skills to provide evidence leading to a conclusion.

For full credit (2 points), students had to identify the cow and describe two of the three characteristics of the hair—texture/smoothness/scales, thickness/size, or color/pattern. Students correctly identifying the cow but describing only one of the characteristics were awarded partial credit (1 point).



13 Farm Investigation: Microscope Pictures

A. Look at the microscope pictures above. Which animal's hair matches the hair from the garden?

(Click one box.)

☑ cow

□ goat

B. Describe **two** things from the microscope pictures above that helped you make your choice.

1.

similar size

2.

same color

Maximum Score Points: 2
Content Domain: Life Science
Topic Area: Characteristics and Life Processes of Organisms
Cognitive Domain: Reasoning
Science Practices: Working with Data,
Answering the Research Question,
Making an Argument from Evidence

As shown in Exhibit 25, only 21 percent of students, on average internationally, achieved full credit by correctly selecting the cow and providing two correct characteristics of the hair. However, a further 31 percent achieved partial credit by selecting the cow and providing just one characteristic. Girls performed better than boys (24% vs. 19%, on average internationally).

Exhibit 25

Farm Investigation Screen 13 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Selects Cow and Gives 2 Correct Characteristics)			Percent Partial Credit (Selects Cow but Gives Only 1 Correct Characteristic)	Percent Selects Cow but No Correct Characteristics (Incorrect)
	Overall Country	Girls	Boys		
Korea, Rep. of	46 (2.7)	50 (3.9)	43 (3.8)	30 (2.5)	19 (2.1)
England	35 (2.5)	38 (3.9)	32 (3.6)	38 (2.4)	20 (1.8)
Singapore	30 (2.0)	34 (2.6)	27 (2.7)	38 (2.0)	26 (1.9)
Norway (5)	30 (2.7)	31 (3.7)	29 (3.5)	37 (2.7)	24 (2.4)
United States	27 (1.8)	32 (2.8)	22 (2.4)	36 (1.7)	30 (1.8)
Malta	26 (2.1)	27 (3.0)	26 (2.8)	29 (1.9)	32 (2.2)
Turkey (5)	25 (2.3)	27 (3.3)	22 (2.8)	24 (2.2)	38 (2.4)
Canada	24 (1.4)	25 (2.1)	23 (2.0)	36 (1.6)	28 (1.4)
Spain	23 (1.9)	24 (2.8)	22 (2.8)	31 (2.5)	34 (2.5)
Croatia	23 (2.2)	26 (3.7)	21 (2.7)	24 (2.2)	39 (3.2)
Denmark	23 (2.2)	27 (2.9)	19 (3.0)	29 (2.2)	40 (2.5)
Portugal	22 (2.2)	22 (3.6)	22 (2.4)	32 (2.9)	32 (2.5)
Finland	22 (2.0)	27 (2.7)	19 (2.8)	34 (2.4)	34 (2.2)
Slovak Republic	22 (2.3)	24 (3.2)	21 (3.2)	30 (2.0)	31 (2.3)
Russian Federation	22 (1.9)	24 (2.5)	20 (2.2)	30 (1.9)	34 (2.2)
Chinese Taipei	22 (2.0)	23 (2.5)	21 (3.0)	39 (2.8)	30 (2.3)
Sweden	21 (2.2)	25 (3.5)	17 (2.9)	32 (2.4)	35 (2.6)
Lithuania	20 (2.2)	24 (3.6)	16 (2.5)	35 (2.4)	30 (2.3)
Netherlands	20 (2.0)	20 (2.9)	19 (2.7)	30 (2.3)	40 (2.7)
Hong Kong SAR	19 (3.3)	22 (4.5)	16 (3.2)	31 (3.1)	43 (2.9)
Czech Republic	18 (1.8)	20 (2.6)	17 (2.5)	33 (2.2)	40 (2.4)
Hungary	17 (1.7)	19 (2.7)	16 (2.1)	34 (2.0)	33 (2.3)
Italy	16 (2.1)	19 (3.4)	13 (2.5)	33 (2.4)	35 (2.6)
Chile	15 (2.1)	19 (3.2)	13 (2.3)	37 (2.5)	31 (2.2)
Austria	14 (1.7)	17 (2.7)	12 (2.1)	31 (2.3)	45 (2.6)
Germany	14 (2.0)	17 (3.0)	10 (2.3)	33 (2.4)	42 (2.6)
Qatar	12 (1.8)	14 (2.1)	10 (2.4)	15 (1.8)	46 (2.2)
United Arab Emirates	12 (0.6)	13 (1.1)	11 (0.8)	21 (0.9)	42 (0.9)
Georgia	9 (1.6)	13 (2.8)	7 (1.9)	13 (2.1)	47 (2.5)
France	8 (1.3)	7 (1.7)	8 (2.3)	24 (2.2)	47 (2.6)
International Average	**21 (0.4)**	**24 (0.6)** ▲	**19 (0.5)**	**31 (0.4)**	**35 (0.4)**
Benchmarking Participants					
Madrid, Spain	28 (2.8)	36 (4.0)	20 (2.7)	33 (2.9)	31 (2.7)
Moscow City, Russian Fed.	27 (1.9)	34 (2.7)	20 (2.7)	30 (2.0)	33 (2.3)
Ontario, Canada	27 (2.3)	28 (3.6)	26 (3.1)	33 (2.6)	29 (2.8)
Quebec, Canada	23 (2.5)	22 (3.3)	24 (3.9)	40 (2.8)	24 (2.3)
Dubai, UAE	19 (1.3)	21 (2.2)	16 (1.8)	33 (1.9)	36 (1.8)
Abu Dhabi, UAE	8 (0.9)	9 (1.3)	8 (1.3)	15 (1.0)	45 (1.7)

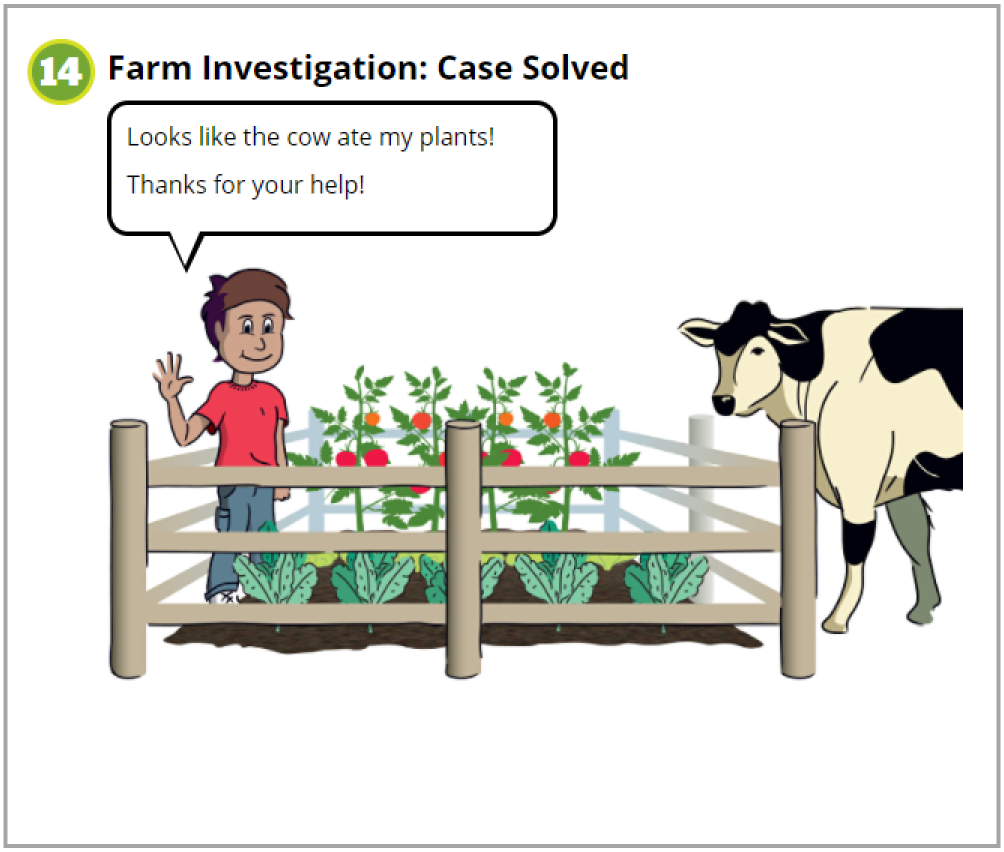
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 14 – Case Solved

On the final screen, George confirms that the cow was the culprit and thanks the students for their help in solving the case. Students who were able to work through the *Farm Investigation* task should have the satisfaction of helping George track down the guilty animal.



14 Farm Investigation: Case Solved

Looks like the cow ate my plants!

Thanks for your help!

Conclusion and Reflections

As an inquiry task that provided fourth grade students with an engaging problem of appropriate difficulty to investigate, the *Farm Investigation* worked very well.

- Students mostly were able to persevere until the end of the task, even when the task was later in the assessment session and students may have been becoming fatigued.
- As a new departure for the TIMSS science assessment, the *Farm Investigation* aimed to capitalize on the ability of the computer to provide a more authentic science inquiry experience than is possible using paper and pencil. Accordingly, although items addressed students' science knowledge and reasoning skills where possible, there was an emphasis on data collection and measurement skills, such as measuring the length of the animal footprints, working through the Animal Finder App to summarize their observations, and using the microscope to examine the animals' hairs.

- Students generally performed well on the items involving the data collection and measurement skills. In fact, they often performed better on these items than on items requiring science knowledge or especially science reasoning skills. However, perhaps because these items did not require any science content knowledge, performance generally showed lower correlations with overall science achievement than more conventional items.
- While future inquiry tasks certainly should include data collection/measurement activities, both for authenticity and for their capacity to engage and motivate, they also should be careful not to neglect items addressing solid science content, to ensure that the task as a whole sufficiently represents the TIMSS Science Assessment Framework.

CHAPTER 3

Mathematics Grade 8

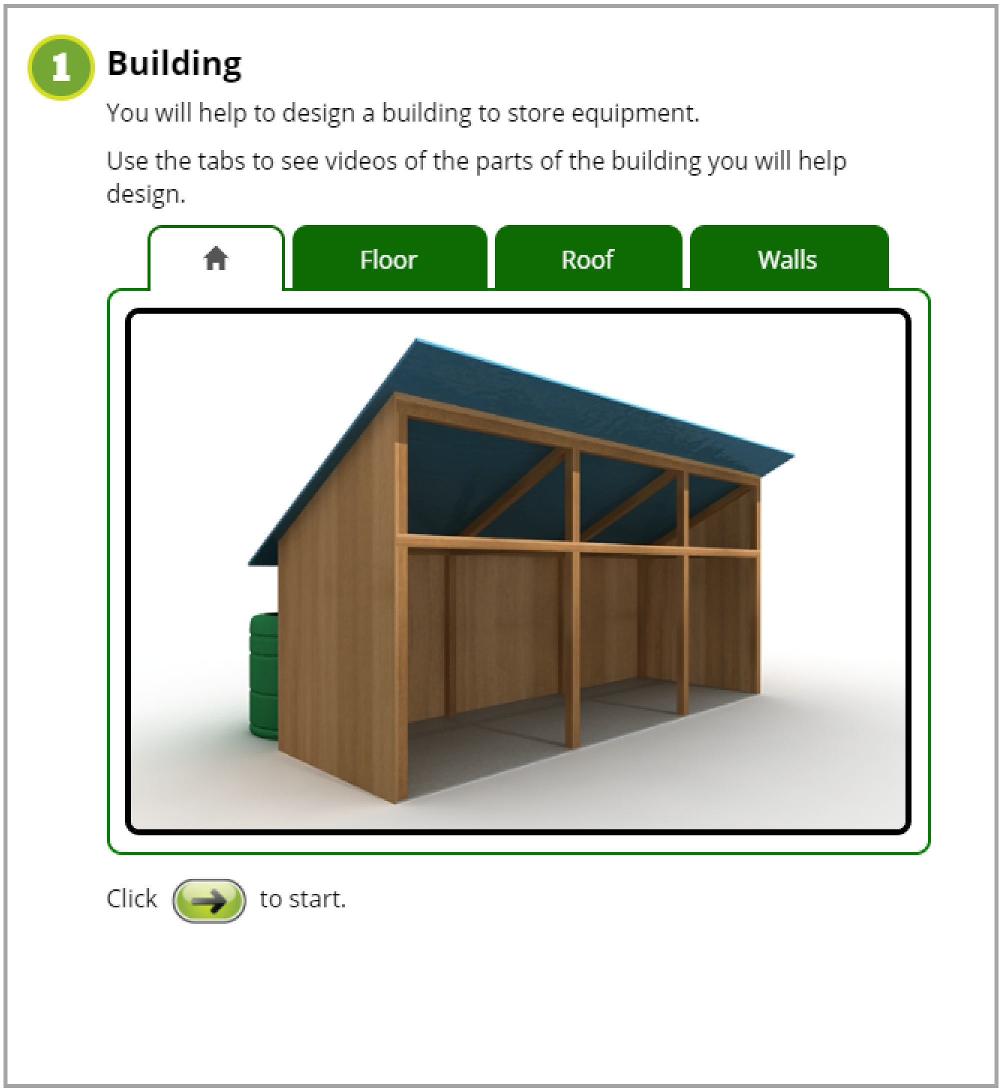
Building

About the Task

The *Building* PSI task assesses eighth grade students' ability to visualize geometric figures and calculate dimensions of lengths, areas, and volumes as they participate virtually in constructing a three-sided storage shed. They did not actually need to "design" the shed as suggested by the introductory text, but they did need to be able to visualize what the shed needed to look like at various stages during the construction process. This PSI task does have a sequential order beginning with the floor, then roof, walls, and tank to catch and store rain. The *Building* PSI task was in a block that also included three shorter items based on robots that helped the students solve algebraic equations. The *Robots* results follow the discussion of the *Building* task.

Screen 1 – Introduction

To help students understand the final goal underlying the sequence of construction activities, the *Building* task begins with a video that shows a revolving 360 degree view of the finished shed, complete with its rain storage tank. As work begins, students can use the tabs (Floor, Roof, Walls) as many times as they like to watch videos showing views of the floor, roof, and walls under construction.



1 Building

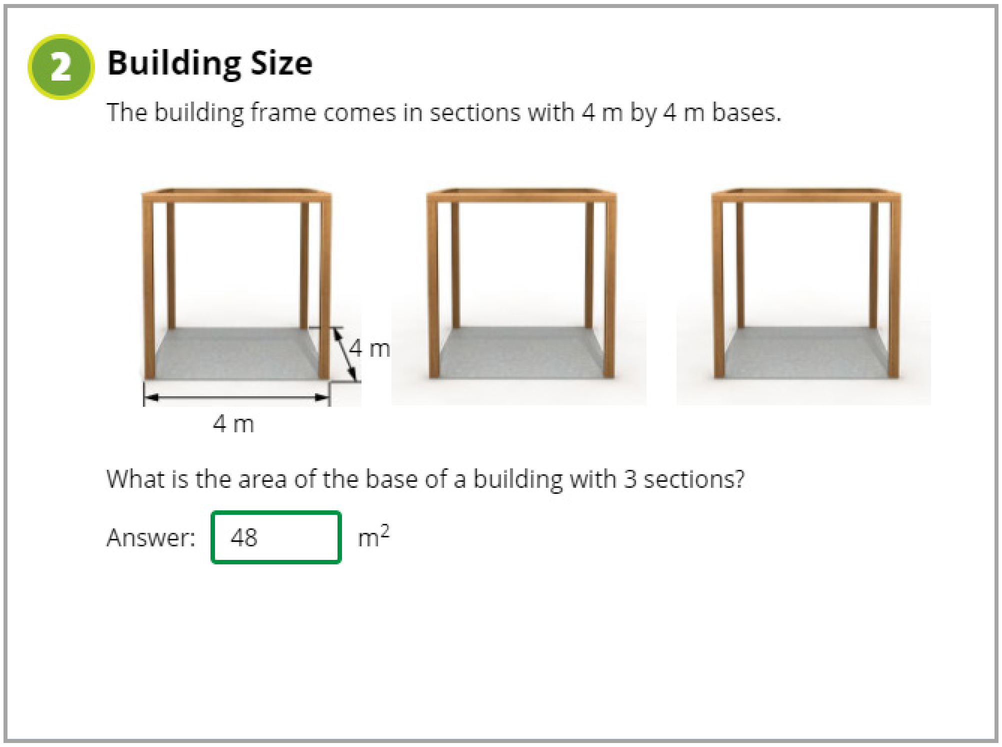
You will help to design a building to store equipment.

Use the tabs to see videos of the parts of the building you will help design.

Click ➔ to start.

Screen 2 – Building Size

Similar to the approach used to begin most PSI tasks, the first item of the *Building* task was intended to be relatively easy for the eighth grade students. In actuality, the *TIMSS 2019 Mathematics Framework* includes determining the area of rectangles under the measurement content domain at the fourth grade.



2 Building Size

The building frame comes in sections with 4 m by 4 m bases.

What is the area of the base of a building with 3 sections?

Answer: 48 m^2

Maximum Score Points: 1
Content Domain: Geometry
Topic Area: Geometric Shapes and Measurement
Cognitive Domain: Applying

Exhibit 26 contains the results confirming that on average across countries most of the eighth grade students (69%) could calculate the area of a rectangle. This includes 44 percent on average that received credit (1 point) for determining the correct area of one section and then multiplying that area by 3. Another 25 percent of the students on average determined the correct area of one section but did not continue and multiply by 3. TIMSS has found that even when the mathematics is straightforward, even eighth grade students often do not carry out the second step of two-step problems. In this particular item, however, it is difficult to understand how students could ignore the picture of the three sections directly above the answer box.

Across countries on average, girls had higher percentages of correct responses than boys (46% vs. 41%). For example, in Singapore with a national result of 70 percent correct, 74 percent of the Singaporean girls answered correctly compared to 65 percent of the boys.

Exhibit 26

Building Screen 2 – Percent Correct Overall and by Gender

Country	Percent Correct (Correct Area for 3 Sections 48 m^2)			Percent Correct Area for Only 1 Section (16 m^2)
	Overall Country	Girls	Boys	
Singapore	70 (2.1)	74 (2.6)	65 (2.8)	22 (1.7)
Korea, Rep. of	62 (2.2)	63 (3.4)	61 (3.3)	19 (1.9)
Sweden	60 (2.4)	65 (3.2)	56 (3.5)	17 (2.1)
Chinese Taipei	57 (1.9)	56 (2.9)	58 (2.6)	30 (1.7)
Hong Kong SAR	55 (2.7)	58 (3.8)	52 (3.7)	31 (2.8)
Russian Federation	54 (2.4)	58 (3.6)	51 (2.6)	23 (1.9)
Norway (9)	53 (2.7)	55 (3.5)	50 (3.6)	23 (2.4)
Lithuania	49 (2.2)	53 (3.0)	46 (3.3)	23 (1.8)
Malaysia	48 (1.5)	50 (2.4)	45 (2.1)	16 (1.3)
Hungary	44 (2.4)	41 (3.3)	47 (2.9)	12 (1.5)
Portugal	43 (3.0)	43 (4.2)	43 (4.0)	26 (2.2)
Finland	42 (2.0)	43 (3.0)	41 (2.6)	34 (2.1)
Israel	42 (2.6)	44 (3.9)	39 (3.4)	24 (2.2)
United States	38 (1.9)	37 (2.8)	40 (2.5)	28 (1.7)
Italy	38 (2.5)	42 (3.6)	34 (2.9)	32 (2.2)
England	36 (2.9)	38 (3.5)	34 (3.7)	31 (2.5)
Turkey	33 (2.2)	39 (3.0)	27 (2.7)	29 (2.0)
United Arab Emirates	29 (1.0)	32 (1.8)	26 (1.5)	25 (1.0)
Georgia	28 (2.9)	29 (4.2)	26 (3.4)	22 (2.6)
Qatar	27 (3.0)	31 (4.3)	23 (3.8)	23 (1.8)
Chile	26 (1.9)	29 (2.6)	22 (2.3)	29 (2.6)
France	25 (1.8)	25 (2.7)	26 (2.7)	33 (2.0)
International Average	**44 (0.5)**	**46 (0.7)** ▲	**41 (0.7)**	**25 (0.4)**
Benchmarking Participants				
Moscow City, Russian Fed.	64 (2.6)	63 (3.1)	65 (3.3)	18 (1.8)
Quebec, Canada	54 (2.6)	59 (3.8)	49 (3.6)	31 (2.7)
Ontario, Canada	53 (3.0)	55 (3.7)	52 (4.2)	23 (2.1)
Dubai, UAE	46 (2.0)	47 (3.2)	44 (3.3)	25 (1.8)
Abu Dhabi, UAE	21 (1.5)	26 (2.8)	16 (1.7)	26 (1.6)

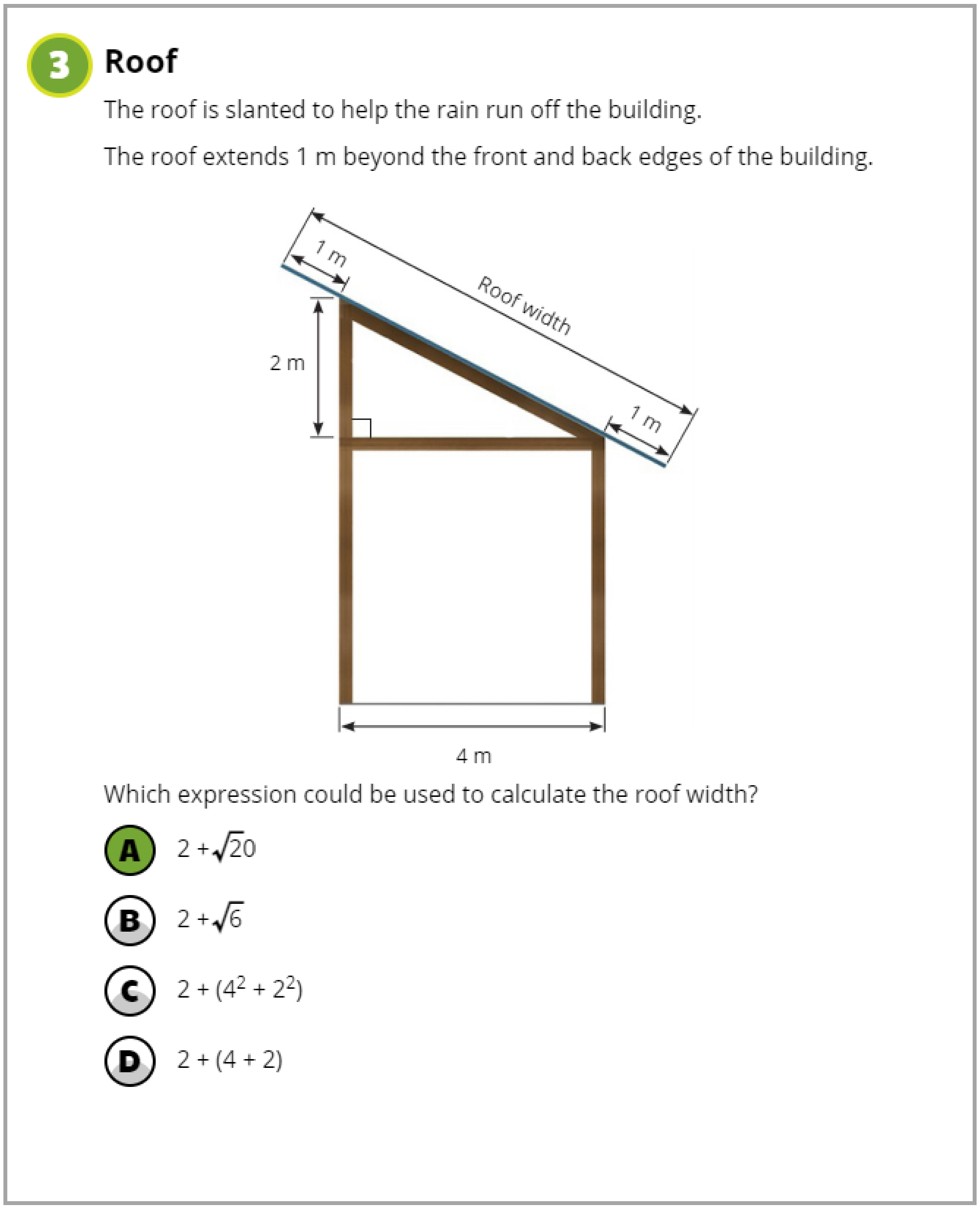
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 3 – Roof

The question about the roof required eighth grade students to select the expression to calculate the width of the roof that was based on the Pythagorean Theorem. Students needed to read the lengths shown in the diagram to determine that the length of the base of the right angle triangle supporting the roof was 4 m. They did not have to actually calculate the width of the roof or necessarily even sum the squares of 2 m and 4 m.



3 Roof

The roof is slanted to help the rain run off the building.

The roof extends 1 m beyond the front and back edges of the building.

Which expression could be used to calculate the roof width?

[✓ A] 2 + $\sqrt{20}$

B 2 + $\sqrt{6}$

C 2 + $(4^2 + 2^2)$

D 2 + (4 + 2)

Maximum Score Points: 1
Content Domain: Geometry
Topic Area: Geometric Shapes and Measurement
Cognitive Domain: Applying

Exhibit 27 shows that, on average across countries, 23 percent of eighth grade students answered this item correctly. Three East Asian countries had the highest achievement—Singapore, Chinese Taipei, and Hong Kong SAR (56, 52, 42 percent correct, respectively). The next highest achievement was by the Russian Federation and Italian students with 27 and 26 percent correct, respectively. There was a slight difference in average achievement favoring boys.

Building Screen 3 – Percent Correct Overall and by Gender

Country	Percent Correct (2 + √20)		
	Overall Country	Girls	Boys
Singapore	56 (2.1)	53 (2.9)	59 (2.5)
Chinese Taipei	52 (2.2)	50 (2.8)	55 (2.9)
Hong Kong SAR	42 (2.9)	38 (3.9)	45 (3.4)
Russian Federation	27 (2.3)	29 (3.9)	25 (2.8)
Italy	26 (2.3)	28 (3.2)	24 (3.1)
Korea, Rep. of	25 (1.8)	22 (2.5)	28 (3.0)
United Arab Emirates	22 (0.9)	19 (1.6)	23 (1.2)
Israel	21 (1.9)	18 (2.8)	23 (2.9)
Portugal	21 (2.2)	21 (3.3)	20 (2.8)
United States	20 (1.8)	19 (2.4)	21 (2.2)
Georgia	20 (2.1)	19 (3.0)	20 (2.8)
Norway (9)	18 (2.0)	19 (2.6)	17 (2.4)
Malaysia	18 (1.3)	18 (1.8)	18 (1.8)
Turkey	17 (1.6)	14 (2.0)	19 (2.3)
Hungary	16 (1.6)	15 (2.3)	17 (2.2)
England	15 (1.6)	17 (2.2)	14 (2.3)
France	15 (1.7)	16 (2.1)	14 (2.2)
Sweden	15 (1.8)	16 (2.9)	14 (2.4)
Lithuania	15 (1.7)	13 (2.2)	17 (2.4)
Chile	14 (1.6)	12 (2.1)	15 (2.2)
Finland	13 (1.4)	12 (1.8)	14 (2.0)
Qatar	13 (1.8)	12 (2.5)	14 (3.0)
International Average	**23 (0.4)**	**22 (0.6)**	**23 (0.5)** ▲
Benchmarking Participants			
Dubai, UAE	31 (1.5)	28 (2.3)	34 (3.0)
Moscow City, Russian Fed.	27 (2.4)	21 (3.3)	32 (3.0)
Ontario, Canada	19 (2.9)	18 (3.5)	20 (3.0)
Abu Dhabi, UAE	16 (1.2)	14 (2.0)	18 (1.7)
Quebec, Canada	12 (1.6)	10 (2.2)	13 (2.0)

▲ Percentage significantly higher than other gender

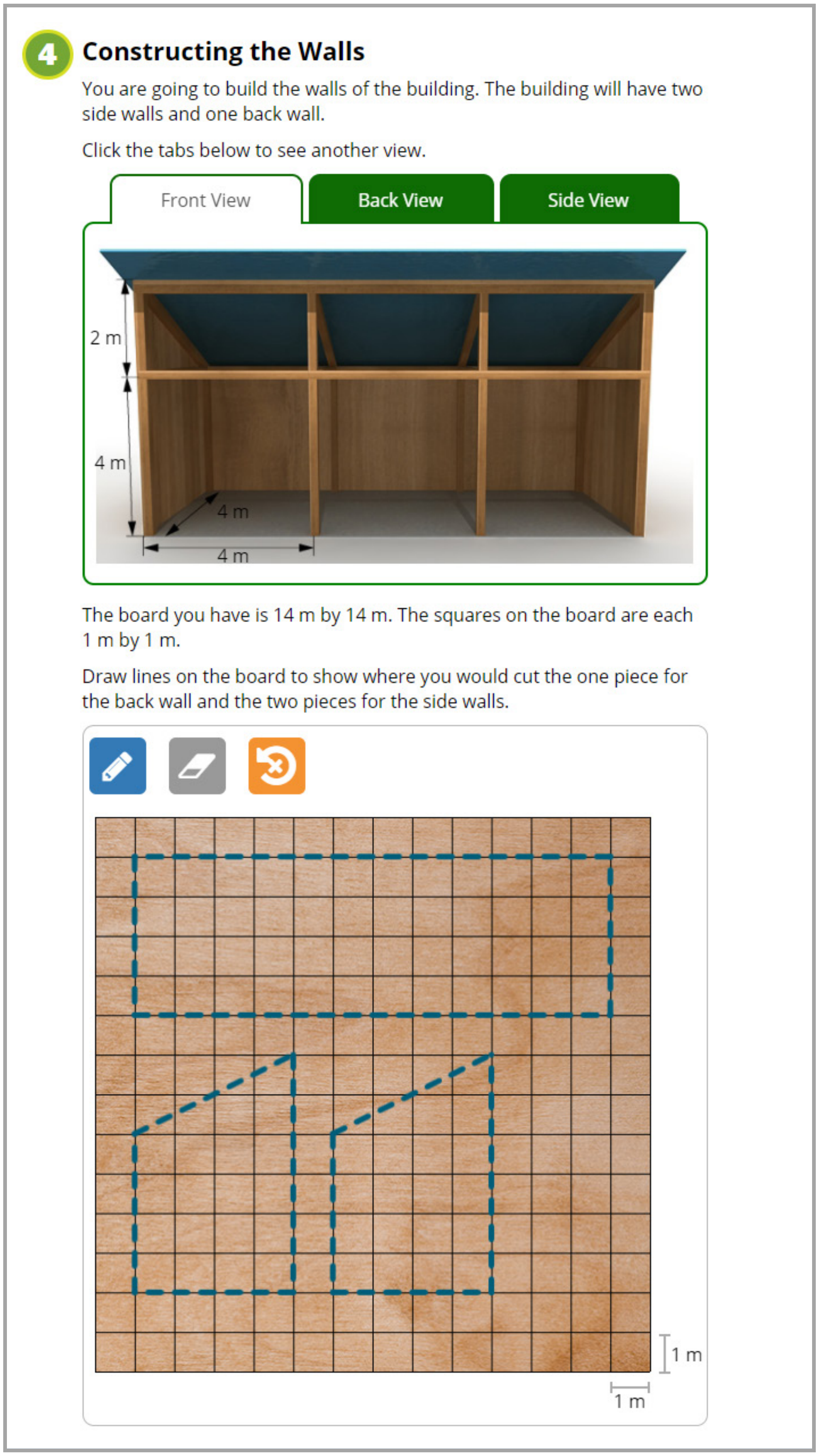
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 4 – Constructing the Walls

This item of the *Building* task involved understanding measurements and reasoning with geometric shapes. Students were presented with a 14 m × 14 m board marked with a grid of 1 m × 1 m squares. They were asked to use the drawing tool to indicate where they would cut out the piece for the back wall of the shed and the two pieces for the side walls. The three pieces fit on the board with extra room leftover, and there was not a requirement to use the smallest amount of the board.

To help them relate the three dimensional representation of the shed to the two-dimensional board, students could use the tabs to view the dimensions of the back of the shed as well as the dimensions of one side and the front of the shed. For full credit (2 points), students needed to fit all three wall pieces—back and two sides—onto the board. The back wall was 4 m × 3 long (for the shed's three sections) by 4 m high. Calculating the dimensions of the sidewalls did require realizing that a side wall was comprised of a 4 m × 4 m square and the right angle triangle holding up the roof which had been the focus of the previous screen (requiring application of the Pythagorean Theorem), and of course, that there were two side walls.



4 Constructing the Walls

You are going to build the walls of the building. The building will have two side walls and one back wall.

Click the tabs below to see another view.

The board you have is 14 m by 14 m. The squares on the board are each 1 m by 1 m.

Draw lines on the board to show where you would cut the one piece for the back wall and the two pieces for the side walls.

Maximum Score Points: 2
Content Domain: Geometry
Topic Area: Geometric Shapes and Measurement
Cognitive Domain: Reasoning

Exhibit 28 shows the percentages of responses awarded full credit (2 points) for correctly measuring and including all three walls. Having already demonstrated some understanding of the dimensions of the right angle triangle that formed the top of each side (in their responses to Screen 3), the students in Singapore, Chinese Taipei, and Hong Kong SAR once again had the highest percentages of full credit responses with 53, 45, and 42 percent fully correct, respectively. However, only 26 percent of the students on average across the eTIMSS countries received full credit (2 points), with another 11 percent of the students on average receiving partial credit (1 point) for fitting the back wall rectangle on the board. There was very little difference in achievement between girls and boys.

Exhibit 28

Building Screen 4 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Draws Back Wall and 2 Sides)			Percent Partial Credit (Draws Back Wall Only)
	Overall Country	Girls	Boys	
Singapore	53 (2.2)	55 (2.7)	52 (2.8)	15 (1.3)
Chinese Taipei	45 (2.4)	46 (2.7)	45 (3.2)	12 (1.5)
Hong Kong SAR	42 (3.1)	47 (4.2)	38 (4.6)	9 (1.5)
Finland	36 (2.0)	35 (2.9)	38 (2.9)	10 (1.3)
Korea, Rep. of	36 (2.4)	35 (2.8)	36 (3.6)	16 (2.0)
Norway (9)	34 (3.0)	37 (4.1)	31 (3.8)	9 (1.3)
Sweden	34 (2.8)	33 (3.7)	34 (3.7)	9 (1.5)
Russian Federation	28 (2.3)	26 (2.8)	30 (3.5)	12 (1.4)
Lithuania	26 (2.4)	28 (3.2)	24 (3.6)	11 (1.8)
United States	26 (2.1)	28 (3.0)	23 (2.5)	12 (1.1)
Portugal	24 (2.4)	22 (2.8)	27 (3.7)	15 (1.8)
England	24 (2.2)	26 (3.6)	23 (3.0)	12 (1.5)
Malaysia	23 (1.6)	23 (2.0)	22 (2.3)	16 (1.1)
Israel	21 (2.2)	26 (3.5)	17 (2.9)	14 (1.7)
Hungary	19 (1.7)	19 (2.1)	20 (2.5)	12 (1.5)
Chile	19 (1.8)	19 (2.4)	18 (2.4)	15 (2.3)
Italy	18 (1.8)	20 (2.6)	15 (2.5)	10 (1.6)
Turkey	18 (2.0)	16 (2.3)	19 (2.5)	12 (1.2)
France	15 (1.7)	15 (2.4)	14 (2.3)	9 (1.5)
United Arab Emirates	14 (0.9)	16 (1.3)	12 (1.1)	9 (0.5)
Qatar	12 (2.2)	13 (2.4)	11 (2.9)	7 (1.3)
Georgia	9 (1.5)	9 (2.2)	8 (2.6)	5 (1.1)
International Average	**26 (0.5)**	**27 (0.6)**	**25 (0.6)**	**11 (0.3)**
Benchmarking Participants				
Moscow City, Russian Fed.	38 (2.4)	34 (3.4)	41 (3.1)	14 (1.7)
Ontario, Canada	34 (2.8)	38 (3.8)	31 (3.8)	13 (1.8)
Quebec, Canada	34 (2.7)	37 (3.3)	31 (3.6)	18 (2.4)
Dubai, UAE	25 (1.8)	28 (2.7)	23 (2.8)	12 (1.2)
Abu Dhabi, UAE	9 (1.3)	12 (2.4)	6 (1.1)	9 (1.1)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

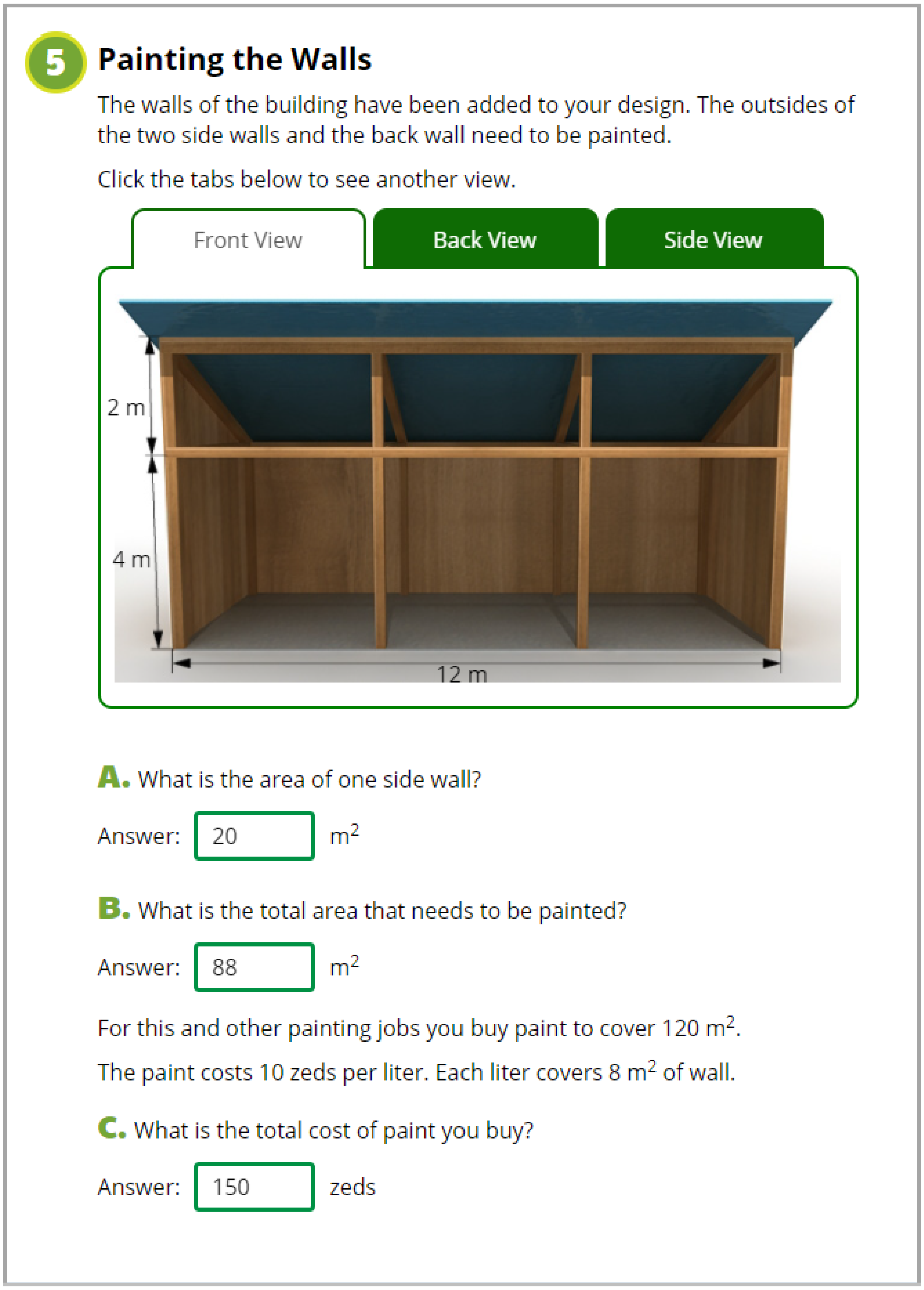
SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Students rarely are presented with a “blank canvas” and asked to proceed on their own. Much more research could be done trying to determine whether the students planned or not, and if so, was there any evidence of a forward-looking strategy. For example, students were awarded partial credit for fitting the back wall onto the board. However, putting that wall in the middle of the board would not be a good strategy for leaving room for the side walls and obtaining two points.

As an example of the potential for using the TIMSS 2019 PSI data for research, the TIMSS & PIRLS International Study Center used the responses from this graphical constructed-response item to explore the possibilities of using automated scoring in future TIMSS assessment cycles (see Appendix C). Artificial neural networks (ANNs) were trained according to the TIMSS scoring guide and example responses to classify the students’ responses. The ANNs scoring was at least as reliable as the human scoring, suggesting that in the future TIMSS might be able to replace the second human scorer in reliability scoring with ANNs scoring.

Screen 5 – Painting the Walls

After having constructed the walls, students’ next task involved painting the walls. Item 5A asked students to calculate the area of one side of the wall. Interestingly, this was the third question in a row that included some understanding of the top triangle of the side wall. Recognizing that the triangle top of the wall was a right angle triangle was integral to solving Screen 3’s question about the width of the roof, with Singapore and Chinese Taipei having the top performance. Next, as part of Screen 4, students needed to determine the dimensions of the side walls (each a top triangle and bottom rectangle) to cut them out of the board, again with Singapore followed by Chinese Taipei and Hong Kong SAR as the top performers. Next, in 5A students were asked to determine the area of one side wall (top triangle plus bottom rectangle).



5 Painting the Walls

The walls of the building have been added to your design. The outsides of the two side walls and the back wall need to be painted.

Click the tabs below to see another view.

A. What is the area of one side wall?

Answer: 20 m^2

B. What is the total area that needs to be painted?

Answer: 88 m^2

For this and other painting jobs you buy paint to cover 120 m^2.

The paint costs 10 zeds per liter. Each liter covers 8 m^2 of wall.

C. What is the total cost of paint you buy?

Answer: 150 zeds

	Item 5A	Item 5B	Item 5C
Maximum Score Points:	1	1	1
Content Domain:	Geometry	Geometry	Algebra
Topic Area:	Geometric Shapes and Measurement	Geometric Shapes and Measurement	Expressions, Operations, and Equations
Cognitive Domain:	Applying	Reasoning	Reasoning

Exhibit 29 presents the percentages of students determining the area of one side wall. Facilitated by their understanding of the Pythagorean Theorem, the eighth grade students in Hong Kong SAR and Chinese Taipei (68% correct) as well as Singapore (66%) and Korea (63%) had the highest performance. After a relatively large gap, Sweden had the next highest percent correct (46%). On average across the eTIMSS countries, 32 percent of the students provided a correct response to the question about the area of one side wall. In considering whether understanding that the triangle was a right angle triangle was the key to success, it is interesting that 15 percent of the students on average treated the top triangle section as if it were another rectangle (making it twice the correct area). There was little or no difference in average achievement between girls and boys.

Exhibit 29

Building Screen 5A – Percent Correct Overall and by Gender

Country	Percent Correct (Correct Area of One Side Wall 20 m^2)			Percent Incorrect Area for Top Triangular Section (24 m^2)
	Overall Country	Girls	Boys	
Hong Kong SAR	68 (2.1)	73 (3.4)	63 (3.0)	8 (1.3)
Chinese Taipei	68 (2.0)	66 (2.7)	69 (2.8)	11 (1.4)
Singapore	66 (2.2)	64 (2.8)	68 (3.0)	9 (1.2)
Korea, Rep. of	63 (2.2)	60 (3.4)	65 (3.1)	11 (1.7)
Sweden	46 (2.5)	45 (3.6)	47 (3.5)	11 (1.4)
Norway (9)	37 (2.6)	35 (3.4)	38 (3.4)	15 (1.9)
England	32 (2.6)	29 (3.1)	36 (3.8)	17 (1.8)
Russian Federation	31 (2.2)	30 (3.5)	31 (3.0)	16 (1.5)
Lithuania	30 (2.2)	33 (3.5)	28 (2.6)	20 (1.9)
Israel	30 (2.3)	34 (3.5)	27 (2.8)	20 (1.7)
Finland	30 (2.2)	31 (3.1)	29 (2.8)	19 (1.5)
Portugal	29 (2.7)	27 (3.3)	32 (4.1)	18 (1.9)
Italy	29 (2.3)	29 (3.5)	28 (2.9)	12 (1.7)
United States	22 (1.7)	19 (2.1)	24 (2.3)	15 (1.3)
Hungary	20 (1.9)	17 (2.5)	24 (2.4)	13 (1.5)
France	20 (2.0)	16 (2.5)	23 (3.0)	16 (1.9)
Turkey	18 (2.0)	16 (2.3)	20 (2.9)	18 (1.9)
United Arab Emirates	17 (0.8)	16 (1.3)	18 (1.1)	11 (0.6)
Malaysia	15 (1.3)	14 (1.9)	15 (1.7)	21 (1.4)
Qatar	13 (2.3)	13 (3.6)	13 (2.6)	10 (1.3)
Georgia	11 (1.9)	10 (2.2)	11 (2.4)	15 (2.3)
Chile	8 (1.3)	9 (2.2)	7 (1.5)	13 (1.5)
International Average	**32 (0.4)**	**31 (0.6)**	**32 (0.6)**	**15 (0.3)**
Benchmarking Participants				
Quebec, Canada	47 (3.0)	43 (3.8)	51 (3.8)	17 (2.0)
Moscow City, Russian Fed.	36 (2.4)	31 (3.5)	40 (3.4)	21 (1.8)
Ontario, Canada	35 (3.4)	36 (3.9)	34 (3.9)	23 (2.5)
Dubai, UAE	32 (1.8)	30 (2.8)	34 (2.6)	14 (1.1)
Abu Dhabi, UAE	9 (0.9)	8 (1.4)	9 (1.3)	11 (1.1)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Item 5B required students to calculate the total area that needed to be painted.

As shown in Exhibit 30, Singapore (48%), Chinese Taipei (43%), and Hong Kong SAR (41%) as well as Korea (40%) had the highest performance. The average percent correct across all countries was considerably lower—20 percent. Interestingly, 7 percent of the students on average used the incorrect area for the side walls, yet still correctly calculated the total area, underscoring the complexities inherent in trying to recover from dependence among items. There were no differences in achievement between girls and boys.

Exhibit 30

Building Screen 5B – Percent Correct Overall and by Gender

Country	Percent Correct Total Area using Correct Area for Side Wall (88 m^2)			Percent Correct Total Area using Incorrect Area for Side Wall
	Overall Country	Girls	Boys	
Singapore	48 (2.1)	49 (2.9)	46 (2.8)	9 (1.3)
Chinese Taipei	43 (2.0)	42 (2.7)	45 (2.8)	5 (0.8)
Hong Kong SAR	41 (2.4)	47 (4.4)	37 (3.5)	5 (1.2)
Korea, Rep. of	40 (2.0)	39 (3.4)	40 (2.9)	5 (0.9)
Sweden	31 (2.5)	33 (3.2)	30 (3.5)	7 (1.0)
Norway (9)	25 (2.0)	25 (2.9)	24 (2.7)	8 (1.1)
Finland	22 (1.9)	24 (2.7)	20 (2.3)	13 (1.4)
Lithuania	21 (2.0)	22 (2.9)	19 (2.5)	11 (1.4)
Russian Federation	20 (1.6)	19 (2.8)	20 (2.7)	9 (1.0)
England	19 (2.0)	18 (2.7)	21 (3.1)	4 (1.0)
Israel	17 (1.8)	16 (2.8)	18 (2.5)	9 (1.4)
Portugal	16 (1.9)	15 (2.5)	18 (2.8)	6 (1.1)
Italy	14 (1.8)	15 (2.9)	13 (2.5)	6 (1.2)
United States	14 (1.6)	13 (2.0)	15 (2.0)	9 (1.0)
Hungary	12 (1.4)	10 (1.8)	15 (2.0)	12 (1.4)
Turkey	10 (1.3)	9 (1.7)	10 (1.8)	8 (1.3)
United Arab Emirates	9 (0.7)	10 (1.3)	9 (0.9)	6 (0.5)
France	9 (1.3)	8 (1.7)	9 (1.8)	9 (1.3)
Malaysia	8 (0.9)	8 (1.3)	8 (1.2)	7 (0.8)
Qatar	7 (1.8)	8 (2.6)	6 (2.3)	3 (0.7)
Georgia	4 (1.3)	4 (1.3)	5 (1.9)	6 (1.6)
Chile	4 (1.3)	5 (2.3)	3 (1.1)	5 (0.9)
International Average	**20 (0.4)**	**20 (0.6)**	**20 (0.5)**	**7 (0.2)**
Benchmarking Participants				
Quebec, Canada	26 (2.5)	28 (3.0)	25 (3.2)	13 (1.7)
Moscow City, Russian Fed.	26 (2.3)	22 (3.0)	29 (3.1)	13 (1.3)
Ontario, Canada	23 (2.9)	24 (3.6)	22 (3.2)	12 (1.7)
Dubai, UAE	20 (1.5)	19 (2.9)	21 (2.3)	8 (1.2)
Abu Dhabi, UAE	4 (0.7)	5 (1.2)	4 (0.8)	4 (0.6)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Fortunately, the painting scenario was reset in 5C by introducing the need to buy some extra paint. Because 10 percent of the students omitted this item, some of them may not have wanted to address this complication.

Exhibit 31 provides the percentages of students providing a correct response. In Chinese Taipei and Singapore, 60–62 percent of the students provided the correct response. Across the participating countries, 33 percent of students answered correctly on average. Further analysis revealed that 6 percent on average performed accurate calculations, but did not read the question thoroughly and used an incorrect area. Once again boys and girls performed similarly.

Building Screen 5C – Percent Correct Overall and by Gender

Country	Percent Correct (150 zeds)		
	Overall Country	Girls	Boys
Chinese Taipei	62 (2.0)	62 (3.3)	62 (2.8)
Singapore	60 (2.0)	61 (3.0)	59 (2.8)
Hong Kong SAR	44 (2.6)	48 (4.0)	42 (3.8)
Norway (9)	41 (2.6)	40 (3.5)	43 (3.6)
Lithuania	40 (2.3)	42 (3.3)	38 (3.0)
Russian Federation	39 (2.4)	36 (3.5)	41 (2.9)
Korea, Rep. of	34 (2.2)	37 (3.4)	32 (3.3)
Italy	34 (2.4)	37 (3.3)	30 (3.0)
Israel	34 (2.0)	30 (2.7)	36 (3.2)
Sweden	33 (2.1)	38 (3.4)	29 (2.9)
United States	32 (1.9)	29 (2.7)	35 (2.5)
England	32 (2.6)	34 (3.7)	29 (3.4)
Turkey	31 (2.2)	29 (2.6)	32 (3.6)
Finland	30 (1.8)	28 (2.8)	32 (2.7)
Hungary	30 (1.8)	29 (2.6)	30 (2.5)
Portugal	26 (2.2)	27 (3.3)	26 (2.9)
France	24 (2.3)	26 (3.0)	22 (3.0)
Malaysia	22 (1.3)	23 (2.1)	20 (1.9)
United Arab Emirates	20 (0.9)	19 (1.3)	21 (1.3)
Chile	19 (1.6)	19 (2.1)	19 (2.5)
Qatar	17 (2.1)	16 (2.5)	18 (3.1)
Georgia	14 (2.2)	16 (3.3)	12 (2.5)
International Average	**33 (0.4)**	**33 (0.6)**	**32 (0.6)**
Benchmarking Participants			
Moscow City, Russian Fed.	42 (2.8)	38 (3.4)	45 (3.8)
Quebec, Canada	42 (3.1)	41 (4.3)	42 (3.7)
Ontario, Canada	35 (2.3)	36 (3.4)	34 (2.7)
Dubai, UAE	32 (2.4)	29 (2.7)	36 (3.4)
Abu Dhabi, UAE	18 (1.1)	19 (1.6)	16 (1.5)

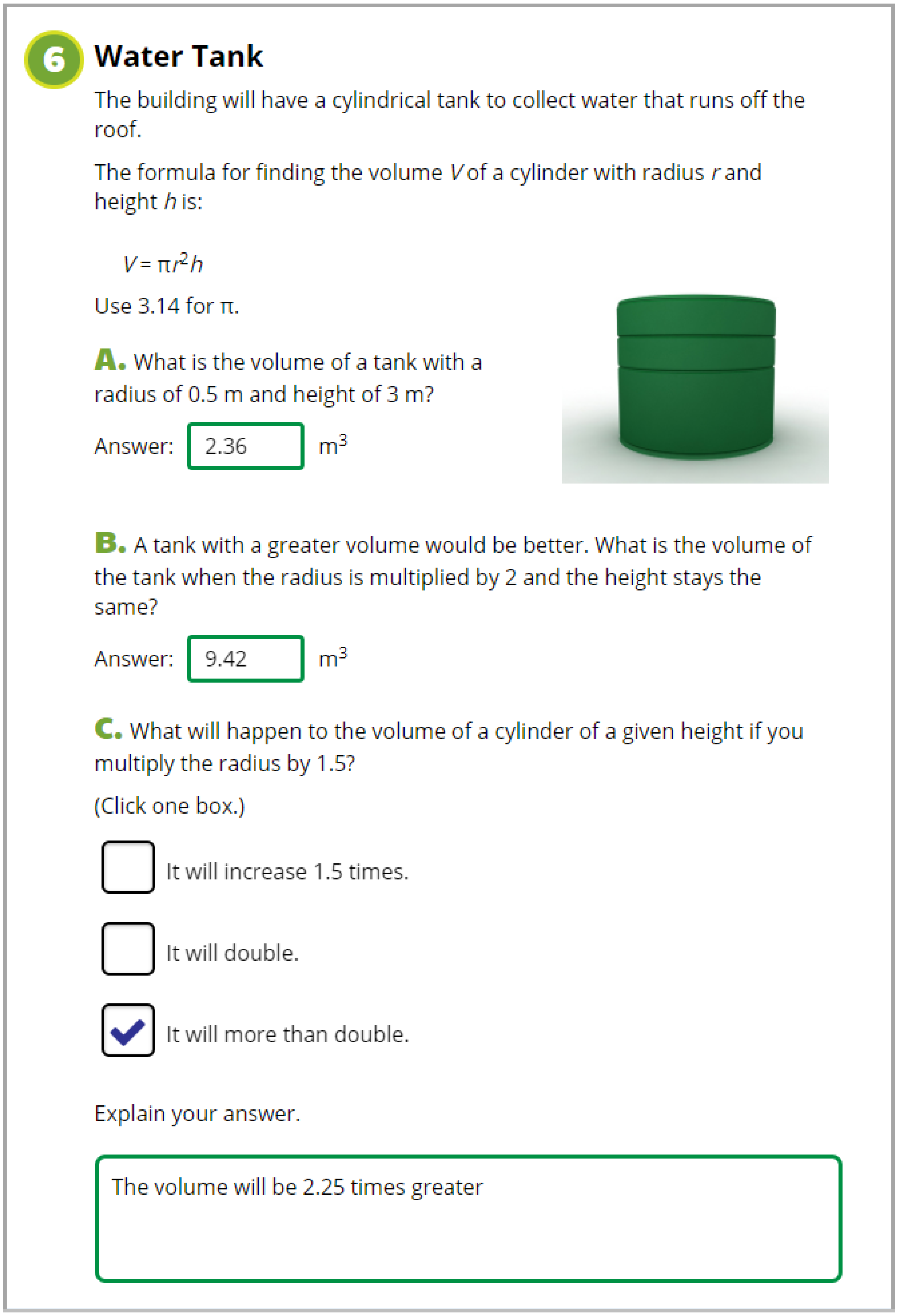
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 6 – Water Tank

Screen 6 shifted the narrative of the scenario from the building to its water tank. Students were given the formula for determining the volume V of a cylinder and asked to use $\pi = 3.14$. However, this may have influenced the 13 percent of the students on average that omitted the items on this screen.

In item 6A, students were asked to calculate the volume of a water tank with a radius of 0.5 m and a height of 3 m. The topic was classified as geometry in the context of this PSI, but the algebra content area also includes: *Find the value of a formula given values of the variables.*



6 Water Tank

The building will have a cylindrical tank to collect water that runs off the roof.

The formula for finding the volume *V* of a cylinder with radius *r* and height *h* is:

$$V = \pi r^2 h$$

Use 3.14 for π.

A. What is the volume of a tank with a radius of 0.5 m and height of 3 m?

Answer: 2.36 m^3

B. A tank with a greater volume would be better. What is the volume of the tank when the radius is multiplied by 2 and the height stays the same?

Answer: 9.42 m^3

C. What will happen to the volume of a cylinder of a given height if you multiply the radius by 1.5?

(Click one box.)

□ It will increase 1.5 times.

□ It will double.

☑ It will more than double.

Explain your answer.

The volume will be 2.25 times greater

	Item 6A	Item 6B	Item 6C
Maximum Score Points:	1	1	2
Content Domain:	Geometry	Geometry	Geometry
Topic Area:	Geometric Shapes and Measurement	Geometric Shapes and Measurement	Geometric Shapes and Measurement
Cognitive Domain:	Applying	Reasoning	Reasoning

Exhibit 32 presents the percentages of students in each country providing the correct answer. About three-fourths of the Singaporean students provided the correct answer, leading the other eTIMSS countries by a substantial margin. The next highest percentages of correct responses represented only about half the students: Hong Kong SAR (52%), Russian Federation (48%), and Chinese Taipei (47%). The average percent correct for the participating countries was 33 percent (mostly due to the 3 countries just mentioned), because 13 countries had achievement of 33 percent correct or lower. Across countries, girls had higher average performance than boys.

Exhibit 32

Building Screen 6A – Percent Correct Overall and by Gender

Country	Percent Correct (Volume of Tank = 2.36 m^3)		
	Overall Country	Girls	Boys
Singapore	76 (2.0)	79 (2.2)	72 (2.9)
Hong Kong SAR	52 (2.7)	55 (3.5)	51 (4.0)
Russian Federation	48 (2.7)	51 (3.5)	46 (3.1)
Chinese Taipei	47 (2.1)	44 (2.7)	50 (2.9)
Korea, Rep. of	43 (1.9)	42 (2.8)	43 (2.9)
Hungary	40 (2.2)	41 (2.8)	40 (2.8)
United States	39 (2.2)	38 (3.0)	40 (2.7)
Lithuania	37 (2.5)	43 (3.3)	30 (3.1)
Sweden	34 (2.5)	35 (3.4)	33 (3.3)
United Arab Emirates	33 (1.1)	38 (1.6)	29 (1.6)
Italy	33 (2.3)	40 (3.3)	24 (2.8)
Norway (9)	32 (2.1)	36 (3.0)	28 (2.9)
England	31 (2.7)	30 (3.3)	31 (3.7)
Portugal	26 (2.0)	26 (2.8)	26 (3.3)
Israel	25 (2.3)	27 (3.0)	24 (3.2)
Qatar	24 (2.1)	25 (3.6)	22 (3.0)
Malaysia	21 (1.3)	24 (2.0)	17 (1.7)
France	20 (1.8)	21 (2.5)	19 (2.4)
Finland	19 (1.9)	21 (2.8)	17 (2.1)
Georgia	16 (2.4)	17 (3.1)	15 (2.9)
Chile	16 (2.4)	18 (4.1)	14 (2.3)
Turkey	15 (2.0)	15 (2.8)	14 (2.0)
International Average	**33 (0.5)**	**35 (0.6)** ▲	**31 (0.6)**
Benchmarking Participants			
Moscow City, Russian Fed.	54 (2.6)	51 (3.8)	56 (3.2)
Dubai, UAE	53 (1.9)	56 (2.8)	51 (3.5)
Quebec, Canada	48 (3.0)	51 (4.1)	47 (3.7)
Ontario, Canada	46 (2.8)	52 (3.7)	41 (3.9)
Abu Dhabi, UAE	24 (1.6)	32 (2.6)	17 (1.5)

▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Item 6B asked students to increase the radius by multiplying it by 2. This essentially removed the radius from the calculation ($0.5 \times 2 = 1$, $1^2 = 1$), leaving 3.14×3.

As shown in Exhibit 33, 5B results closely mirrored those for 5A. Singapore had the highest achievement at 77 percent correct, followed by about half the students in Chinese Taipei (50%), Hong Kong SAR (48%), and Korea (46%). The average across eTIMSS countries was 32 percent correct. On average, there was little or no performance difference between girls and boys.

Building Screen 6B – Percent Correct Overall and by Gender

Country	Percent Correct (Volume of Tank = 9.42 m^3)		
	Overall Country	Girls	Boys
Singapore	77 (2.0)	77 (2.6)	76 (2.7)
Chinese Taipei	50 (2.1)	48 (2.8)	52 (2.7)
Hong Kong SAR	48 (2.7)	50 (3.8)	47 (3.4)
Korea, Rep. of	46 (2.0)	43 (2.7)	49 (3.0)
Russian Federation	46 (2.6)	48 (3.4)	44 (3.1)
United States	38 (2.1)	38 (2.8)	39 (2.6)
Hungary	32 (1.9)	27 (2.6)	37 (2.4)
Lithuania	32 (2.1)	36 (2.8)	27 (3.1)
United Arab Emirates	31 (1.2)	34 (1.6)	27 (1.5)
Italy	30 (2.3)	34 (3.3)	26 (2.7)
Norway (9)	30 (2.2)	34 (3.2)	26 (3.0)
Israel	28 (2.3)	25 (2.7)	30 (3.5)
Sweden	28 (2.5)	28 (3.5)	27 (3.3)
England	27 (2.8)	25 (3.2)	29 (3.8)
France	24 (1.8)	23 (2.6)	24 (2.3)
Portugal	24 (2.1)	24 (3.0)	23 (3.4)
Qatar	23 (2.5)	25 (3.6)	21 (3.2)
Malaysia	22 (1.7)	25 (2.4)	19 (1.8)
Finland	20 (1.8)	23 (2.8)	18 (2.1)
Georgia	14 (2.1)	14 (2.5)	15 (3.0)
Turkey	13 (1.7)	13 (2.0)	13 (2.6)
Chile	13 (1.7)	12 (2.1)	13 (2.2)
International Average	**32 (0.5)**	**32 (0.6)**	**31 (0.6)**
Benchmarking Participants			
Moscow City, Russian Fed.	54 (2.7)	53 (3.7)	55 (3.4)
Dubai, UAE	48 (2.5)	49 (2.7)	47 (4.1)
Ontario, Canada	46 (2.6)	47 (4.0)	45 (3.6)
Quebec, Canada	40 (2.7)	42 (3.7)	39 (3.8)
Abu Dhabi, UAE	24 (1.2)	28 (2.0)	20 (1.5)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Exhibits 34 and 35 provide data about calculator use on 6A and 6B, respectively. On the whole, TIMSS mathematics items are developed to be calculator neutral, such that there is not an advantage or disadvantage to having a calculator. Students were not permitted to bring their own calculators, but the eighth grade students were provided a TIMSS on-screen calculator as part of the eTIMSS user interface. The calculator has the four basic functions, a square root key, and the negative sign. On average more students providing correct answers used the calculator than did not (29% vs. 3% on 6A and 26% vs. 5% on 6B). However, the percentages of students that used calculators to obtain correct answers was similar to the percentages of students that used calculators to obtain incorrect answers (28% for 6A and 24% for 6B). It appears that if the students did know the correct procedure to solve the problems in 6A and 6B, the calculator was a useful tool. However, if the students did not know the procedures for solving the problems, the calculator did not help.

Exhibit 34

Building Screen 6A – Percent of Students Who Used Calculator for Correct and Incorrect Responses

Country	Item Percent Correct	Correct Response		Incorrect Response	
		Used Calculator	Did Not Use Calculator	Used Calculator	Did Not Use Calculator
Singapore	75 (2.1)	74 (0.7)	2 (0.7)	18 (3.1)	4 (3.1)
Hong Kong SAR	52 (2.7)	49 (1.9)	3 (1.9)	27 (3.4)	16 (3.4)
Russian Federation	48 (2.8)	40 (2.6)	7 (2.6)	24 (3.7)	15 (3.7)
Chinese Taipei	47 (2.0)	38 (2.6)	9 (2.6)	16 (2.9)	33 (2.9)
Korea, Rep. of	42 (1.9)	36 (2.3)	6 (2.3)	16 (2.9)	37 (2.9)
Hungary	40 (2.3)	37 (1.6)	2 (1.6)	20 (3.6)	27 (3.6)
United States	38 (2.2)	35 (2.0)	3 (2.0)	36 (2.5)	20 (2.5)
Lithuania	36 (2.4)	28 (4.0)	8 (4.0)	28 (3.6)	25 (3.6)
United Arab Emirates	33 (1.1)	26 (2.0)	7 (2.0)	29 (1.5)	32 (1.5)
Sweden	32 (2.4)	32 (0.8)	0 (0.8)	30 (3.4)	18 (3.4)
Italy	32 (2.3)	30 (1.8)	2 (1.8)	32 (2.9)	17 (2.9)
Norway (9)	30 (2.1)	20 (4.8)	10 (4.8)	13 (3.1)	31 (3.1)
England	30 (2.7)	30 (0.4)	0 (0.4)	29 (3.7)	19 (3.7)
Portugal	26 (2.0)	25 (1.4)	1 (1.4)	37 (2.8)	17 (2.8)
Israel	25 (2.2)	24 (0.8)	0 (0.8)	32 (2.9)	23 (2.9)
Qatar	23 (2.1)	20 (5.0)	3 (5.0)	27 (2.5)	42 (2.5)
Malaysia	21 (1.3)	20 (0.9)	1 (0.9)	52 (2.2)	24 (2.2)
France	20 (1.8)	20 (0.0)	0 (0.0)	39 (2.8)	19 (2.8)
Finland	19 (1.9)	19 (0.8)	0 (0.8)	34 (2.4)	30 (2.4)
Chile	15 (2.3)	15 (1.3)	0 (1.3)	27 (3.7)	33 (3.7)
Georgia	15 (2.3)	13 (5.4)	2 (5.4)	17 (3.6)	35 (3.6)
Turkey	15 (2.0)	12 (4.2)	3 (4.2)	19 (2.3)	52 (2.3)
International Average	**33 (0.5)**	**29 (0.6)**	**3 (0.6)**	**28 (0.6)**	**25 (0.6)**
Benchmarking Participants					
Dubai, UAE	53 (1.9)	45 (2.3)	8 (2.3)	25 (3.0)	17 (3.0)
Moscow City, Russian Fed.	53 (2.6)	52 (0.8)	1 (0.8)	28 (3.5)	9 (3.5)
Quebec, Canada	48 (2.9)	44 (2.5)	4 (2.5)	32 (3.2)	12 (3.2)
Ontario, Canada	45 (2.8)	41 (2.9)	4 (2.9)	33 (3.8)	13 (3.8)
Abu Dhabi, UAE	24 (1.6)	18 (3.6)	7 (3.6)	29 (1.5)	40 (1.5)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Exhibit 35

Building Screen 6B – Percent of Students Who Used Calculator for Correct and Incorrect Responses

Country	Item Percent Correct	Correct Response		Incorrect Response	
		Used Calculator	Did Not Use Calculator	Used Calculator	Did Not Use Calculator
Singapore	76 (2.0)	74 (0.7)	2 (0.7)	15 (3.3)	6 (3.3)
Chinese Taipei	49 (2.1)	33 (2.8)	16 (2.8)	15 (3.0)	31 (3.0)
Hong Kong SAR	48 (2.7)	43 (1.9)	4 (1.9)	26 (3.9)	20 (3.9)
Korea, Rep. of	45 (2.0)	34 (2.4)	12 (2.4)	14 (2.9)	35 (2.9)
Russian Federation	45 (2.7)	38 (2.8)	7 (2.8)	22 (3.3)	19 (3.3)
United States	37 (2.1)	33 (2.0)	4 (2.0)	33 (2.7)	23 (2.7)
Hungary	32 (1.9)	29 (1.9)	3 (1.9)	23 (3.5)	28 (3.5)
Lithuania	31 (2.1)	25 (4.5)	6 (4.5)	28 (3.2)	26 (3.2)
United Arab Emirates	30 (1.2)	22 (2.0)	8 (2.0)	26 (1.7)	35 (1.7)
Italy	29 (2.2)	27 (2.5)	2 (2.5)	29 (3.1)	20 (3.1)
Norway (9)	27 (2.0)	17 (4.1)	10 (4.1)	12 (3.4)	30 (3.4)
Israel	27 (2.3)	25 (2.1)	2 (2.1)	21 (3.8)	28 (3.8)
Sweden	26 (2.4)	25 (1.3)	1 (1.3)	26 (3.0)	23 (3.0)
England	26 (2.7)	25 (1.7)	1 (1.7)	23 (3.7)	24 (3.7)
Portugal	23 (2.1)	22 (1.7)	1 (1.7)	34 (3.2)	22 (3.2)
France	23 (1.7)	22 (1.4)	1 (1.4)	26 (3.1)	23 (3.1)
Qatar	23 (2.4)	17 (4.9)	5 (4.9)	24 (2.7)	40 (2.7)
Malaysia	22 (1.7)	21 (1.6)	1 (1.6)	51 (2.3)	22 (2.3)
Finland	20 (1.8)	19 (1.1)	0 (1.1)	26 (2.5)	33 (2.5)
Georgia	13 (2.0)	9 (7.0)	4 (7.0)	17 (4.0)	35 (4.0)
Turkey	13 (1.7)	8 (6.7)	5 (6.7)	19 (2.4)	51 (2.4)
Chile	12 (1.6)	11 (3.2)	1 (3.2)	23 (3.3)	33 (3.3)
International Average	**31 (0.4)**	**26 (0.7)**	**5 (0.7)**	**24 (0.7)**	**27 (0.7)**
Benchmarking Participants					
Moscow City, Russian Fed.	53 (2.7)	50 (1.3)	3 (1.3)	25 (3.5)	11 (3.5)
Dubai, UAE	48 (2.5)	38 (2.4)	9 (2.4)	25 (3.0)	20 (3.0)
Ontario, Canada	45 (2.5)	40 (3.1)	5 (3.1)	31 (4.0)	14 (4.0)
Quebec, Canada	39 (2.6)	35 (3.6)	4 (3.6)	34 (3.6)	17 (3.6)
Abu Dhabi, UAE	23 (1.2)	16 (3.8)	8 (3.8)	24 (2.0)	43 (2.0)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Perhaps because the problem did not ask for any specific calculations, item 6C about how increasing the radius would impact the volume of the rain barrel left most of the students perplexed. Based on the formula and supported by the pattern of results in 6A and 6B, students could reason that the volume will increase. Students who explained why the volume would more than double received full credit (2 points).

Exhibit 36 presents the results. Only 4 countries managed to achieve at least double-digit success in answering the question for full credit—Singapore and Chinese Taipei (both 13%), Korea (11%), and Hong Kong SAR (10%). The cross-country average was 4 percent. However, another 3 percent on average received partial credit (1 point) for working out an example comparing two cylinders, with one cylinder having 1.5 times the radius of the other. On average across countries, boys slightly outperformed girls (4% vs. 3%).

Building Screen 6C – Percent Full Credit Overall and by Gender

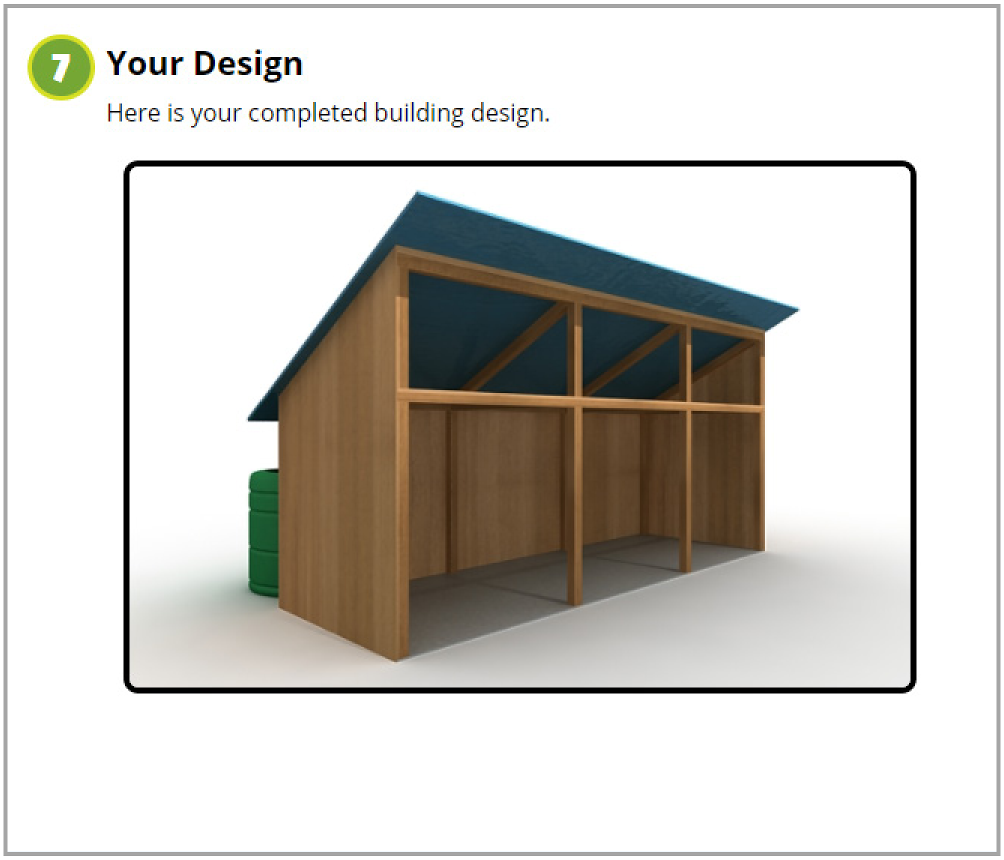
Country	Percent Full Credit (Explains Volume will be 2.25 times Greater)			Percent Partial Credit (Compares Volume of 2 Cylinders)
	Overall Country	Girls	Boys	
Singapore	13 (1.4)	9 (1.5)	16 (2.2)	9 (1.2)
Chinese Taipei	13 (1.7)	9 (1.8)	16 (2.2)	8 (1.1)
Korea, Rep. of	11 (1.5)	9 (1.9)	13 (2.3)	4 (0.9)
Hong Kong SAR	10 (1.8)	7 (2.4)	13 (2.7)	5 (1.1)
Israel	4 (1.0)	4 (1.2)	5 (1.4)	1 (0.5)
Russian Federation	4 (0.9)	3 (1.1)	5 (1.6)	4 (1.0)
Turkey	4 (1.1)	4 (1.6)	3 (0.9)	2 (0.6)
Georgia	2 (0.9)	3 (1.2)	2 (0.9)	0 (0.3)
Hungary	2 (0.6)	2 (0.7)	3 (0.8)	1 (0.5)
United States	2 (0.4)	2 (0.5)	3 (0.7)	4 (0.6)
United Arab Emirates	2 (0.4)	1 (0.3)	3 (0.6)	2 (0.4)
Lithuania	2 (0.6)	1 (0.7)	2 (1.1)	1 (0.4)
Portugal	2 (0.7)	2 (1.1)	1 (0.8)	1 (0.4)
Sweden	2 (0.8)	0 (0.3)	3 (1.4)	4 (1.1)
England	2 (0.6)	1 (0.5)	3 (1.1)	3 (0.7)
Qatar	1 (0.5)	1 (0.5)	1 (0.8)	1 (0.4)
Chile	1 (0.3)	1 (0.4)	1 (0.5)	1 (0.4)
Malaysia	1 (0.3)	1 (0.4)	1 (0.3)	2 (0.4)
France	1 (0.3)	1 (0.4)	1 (0.5)	1 (0.5)
Norway (9)	1 (0.3)	0 (0.1)	1 (0.5)	3 (0.9)
Italy	0 (0.1)	0 (0.0)	0 (0.3)	1 (0.5)
Finland	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.5)
International Average	**4 (0.2)**	**3 (0.2)**	**4 (0.3)** ▲	**3 (0.2)**
Benchmarking Participants				
Moscow City, Russian Fed.	8 (1.5)	5 (1.6)	10 (2.5)	4 (1.0)
Dubai, UAE	4 (1.2)	2 (0.7)	6 (2.0)	3 (0.9)
Ontario, Canada	3 (1.3)	6 (2.9)	1 (0.5)	3 (1.3)
Quebec, Canada	2 (0.9)	0 (0.5)	4 (1.5)	1 (0.5)
Abu Dhabi, UAE	1 (0.3)	1 (0.4)	1 (0.5)	1 (0.2)

▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 7 – Your Design


7 Your Design

Here is your completed building design.

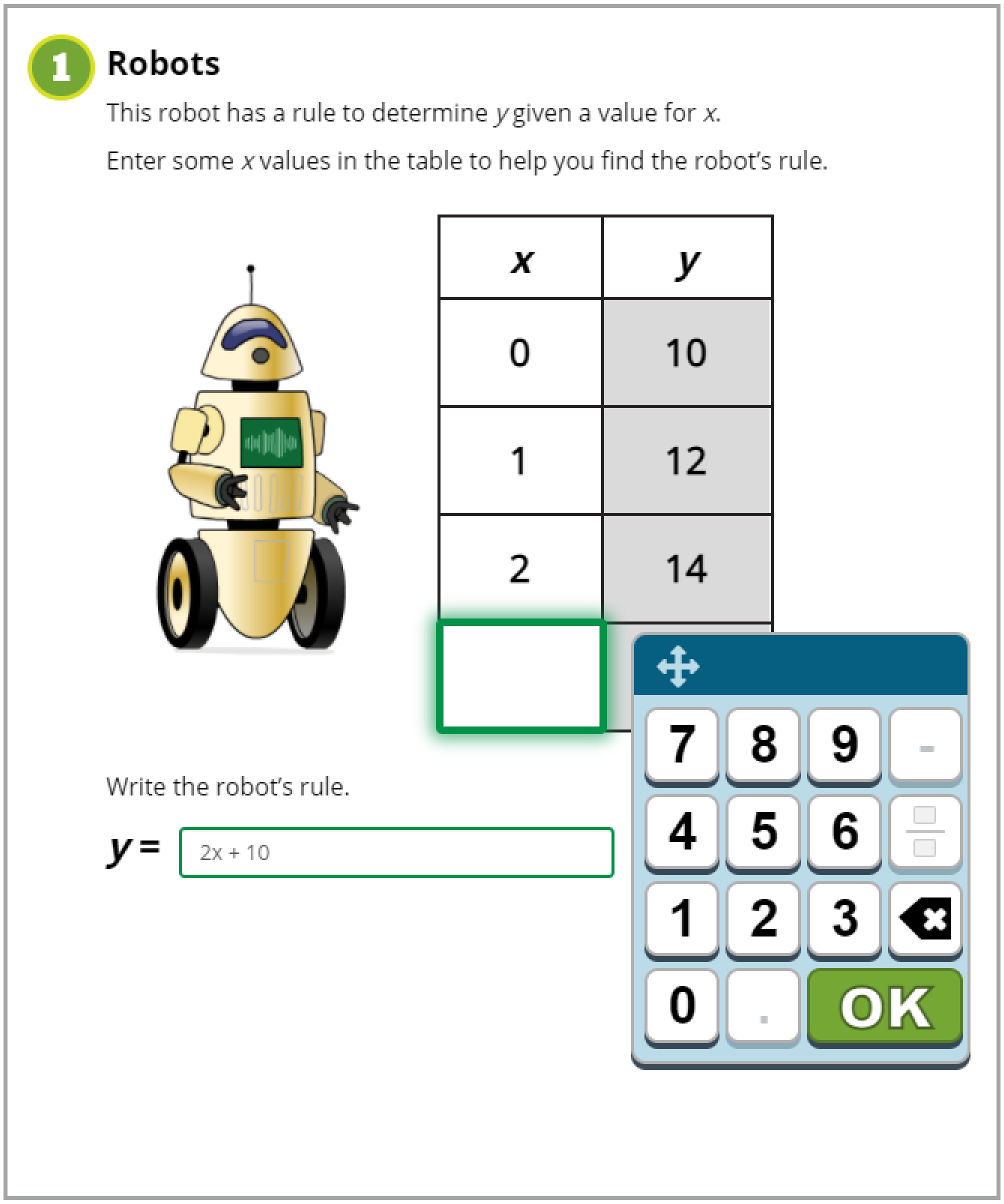
Robots Items

The idea for *Robots* was part of the work to develop PSI tasks, but as the work evolved the TIMSS Robots did not seem to have the necessary requirements to become PSI tasks. Still the idea has potential for future digital assessments. As shown in this section, the TIMSS Robots were programmed to provide the value of *y* for any value of *x*, illustrating a particular type of digital assessment item. The first robot has only one item and the second robot has two items.

The *Robots* results for eighth grade are presented here, because the two item screens followed directly after *Building* in the assessment sessions. It seems that the eighth grade students found the combination of *Building* (geometry/algebra) followed by *Robots* (algebra) challenging. First, 15 percent of the students omitted Screen 1 of *Robots* on average, and then the average percent not reached rose from 3 percent for the first item to 11 percent for the last of the three.

Screen 1

The robot on Screen 1 provided a *y* value for each *x* value submitted by the students. The students were asked to use the number pad to enter some *x* values to determine the robot's rule for determining *y*.



1 **Robots**

This robot has a rule to determine *y* given a value for *x*.

Enter some *x* values in the table to help you find the robot's rule.

x	*y*
0	10
1	12
2	14

Write the robot's rule.

$y =$ 2x + 10

Maximum Score Points: 1
Content Domain: Algebra
Topic Area: Relationships and Functions
Cognitive Domain: Reasoning

As shown in Exhibit 37, 20 percent of the students on average correctly identified the robot's rule (1 point), which was $y = 2x + 10$. Korea (40%), Singapore (39%), and Chinese Taipei (37%) had the highest percentages of eighth grade students identifying the rule. Looking at the process data revealed that the most popular input strategy was entering sequential numbers, although some students entered multiples (2s, 5s, or 10s). On average across countries, boys had a higher percentage correct than girls.

Exhibit 37

Robots Screen 1 – Percent Correct Overall and by Gender

Country	Percent Correct (2*x* + 10)		
	Overall Country	Girls	Boys
Korea, Rep. of	40 (2.5)	36 (3.0)	43 (3.5)
Singapore	39 (2.2)	35 (2.8)	43 (2.7)
Chinese Taipei	37 (2.0)	31 (2.5)	42 (3.1)
Hong Kong SAR	33 (2.5)	30 (4.5)	35 (3.2)
Russian Federation	29 (3.1)	28 (3.8)	31 (3.9)
Turkey	28 (2.1)	28 (2.7)	26 (3.4)
Hungary	25 (1.9)	23 (2.8)	28 (2.8)
Israel	23 (2.1)	22 (2.9)	23 (2.8)
United States	22 (1.9)	19 (2.3)	26 (2.2)
England	20 (2.1)	19 (2.8)	22 (3.4)
United Arab Emirates	19 (0.9)	16 (1.3)	20 (1.3)
Lithuania	17 (1.7)	17 (2.4)	18 (2.4)
Norway (9)	16 (2.0)	14 (2.8)	19 (2.6)
Sweden	16 (1.9)	13 (2.1)	19 (2.9)
Portugal	15 (1.7)	12 (2.0)	18 (3.0)
Finland	14 (1.2)	13 (1.8)	14 (1.8)
Italy	12 (1.5)	12 (2.2)	13 (2.1)
Georgia	11 (1.6)	12 (2.2)	11 (2.1)
France	11 (1.3)	10 (1.7)	13 (2.1)
Qatar	10 (1.4)	9 (1.7)	11 (2.1)
Chile	6 (0.9)	6 (1.7)	6 (1.1)
Malaysia	4 (0.6)	4 (0.9)	3 (0.7)
International Average	**20 (0.4)**	**19 (0.5)**	**22 (0.6)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	39 (2.3)	32 (3.0)	44 (3.4)
Dubai, UAE	33 (2.0)	30 (2.8)	36 (3.0)
Ontario, Canada	24 (2.3)	26 (3.3)	23 (2.7)
Quebec, Canada	21 (2.6)	14 (3.0)	26 (4.0)
Abu Dhabi, UAE	11 (0.9)	11 (1.7)	11 (1.4)

▲ Percentage significantly higher than other gender

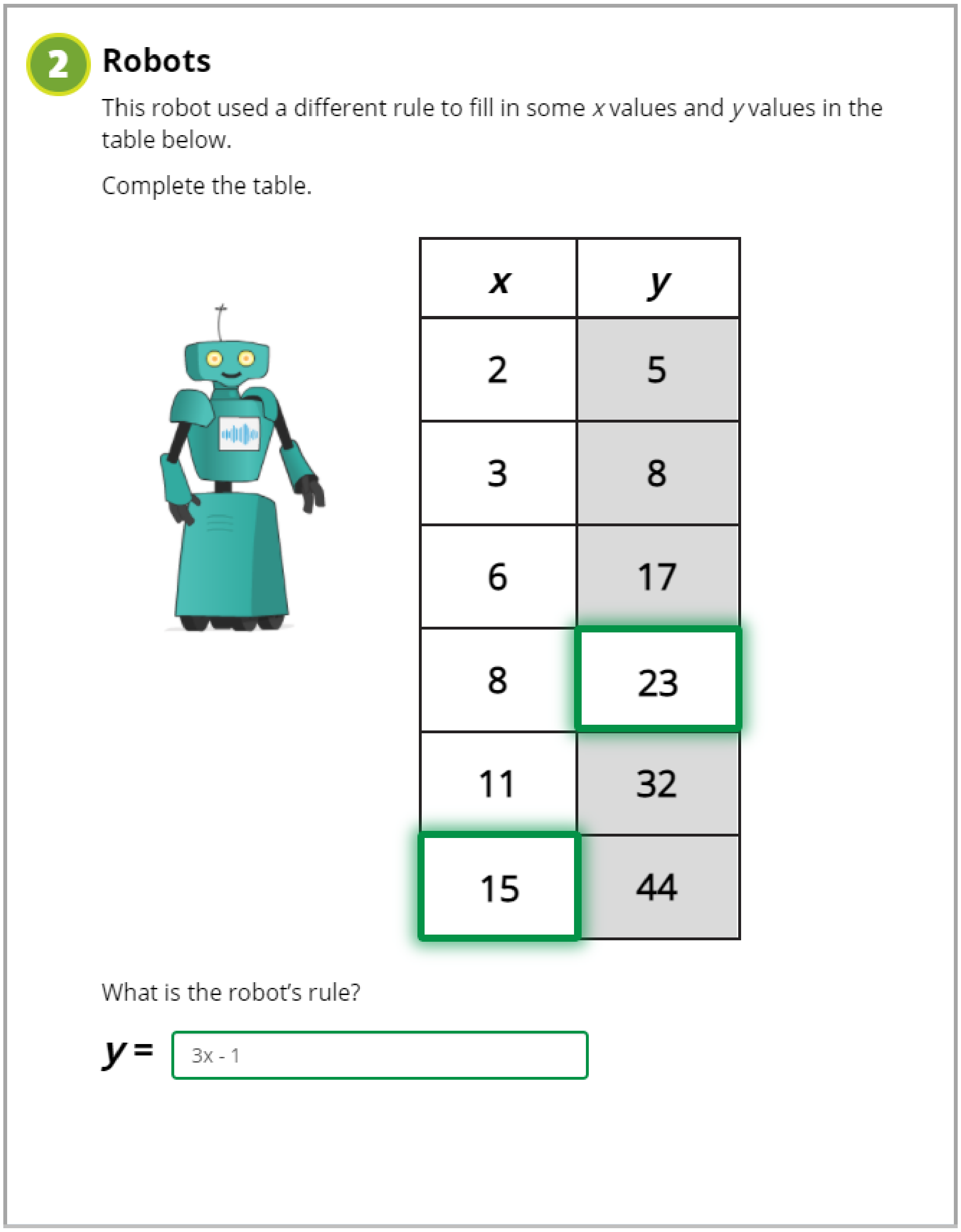
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

As shown above on Screen 1, the table provided to the students included four rows for the students to accommodate a total of four entries, and then permitted reusing the boxes. Although further analysis showed that 53 percent of the students on average realized this and reused the boxes, looking at the process data suggested that some students were less familiar with these types of response spaces. Students' uncertainty about how to use the boxes may have contributed to the omit rate, since 8 percent of the students on average did not make even one entry (another 1 percent made only one entry). Running out of entry boxes may also have had adversely affected achievement, because on average, 37 percent of the students made exactly four entries.

Screen 2

The second robot on Screen 2 had a different rule for determining the relationship between *x* and *y*. The relationship was somewhat more complicated, so the robot made a table of 6 pairs of *x* and *y*, but left out the *y* for the fourth pair and the *x* for sixth pair. The students were asked to complete the table and provide the robot's rule. The non-response rates for Screen 2 were similar to those for Screen 1, but the omit rates increased with 16 percent omitting part A. Again, the format most likely was a contributing factor. This item was not interactive. That is, based on the robot in Screen 1, students may have thought they could make additional entries into the table. But for the second robot, students needed to do their work on separate "scratch" paper.



2 Robots

This robot used a different rule to fill in some *x* values and *y* values in the table below.

Complete the table.

x	*y*
2	5
3	8
6	17
8	23
11	32
15	44

What is the robot's rule?

y = 3x - 1

	Item 2A	Item 2B
Maximum Score Points:	2	1
Content Domain:	Algebra	Algebra
Topic Area:	Relationships and Functions	Relationships and Functions
Cognitive Domain:	Applying	Reasoning

Exhibit 38 contains the results for the percent of students receiving full credit (2 points) or partial credit (1 point) in 2A for finding either *x* or *y*, but not both. Eighth grade students in Chinese Taipei, Singapore, and Korea had the highest achievement—49, 46, and 45 percent of students were awarded full credit, respectively. However, there was a wide range in performance with 6 countries having less than 20 percent fully correct, such that the overall average for fully credit was 25 percent. Another 13 percent on average received partial credit (1 point)—7 percent for finding *y* but not *x* and 6 percent for finding *x* but not *y*. Boys had higher achievement than girls on average across countries.

Exhibit 38

Robots Screen 2A – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Correct Entries for *y* and *x*)			Percent Partial Credit (Only 1 Correct Entry)	
	Overall Country	Girls	Boys	Correct Entry for *y* Only	Correct Entry for *x* Only
Chinese Taipei	49 (2.1)	43 (2.9)	54 (2.8)	7 (1.0)	5 (1.0)
Singapore	46 (2.1)	41 (2.3)	51 (2.9)	8 (1.2)	8 (1.0)
Korea, Rep. of	45 (2.5)	38 (3.3)	51 (3.4)	6 (1.1)	4 (0.8)
Hong Kong SAR	36 (2.4)	35 (3.5)	36 (3.4)	9 (1.5)	7 (1.3)
Israel	32 (2.8)	30 (4.1)	34 (3.7)	6 (1.3)	5 (0.9)
United States	31 (1.8)	27 (2.2)	35 (2.5)	7 (0.8)	5 (0.9)
Hungary	29 (2.0)	25 (2.8)	33 (2.8)	6 (0.9)	3 (0.8)
Turkey	26 (2.4)	25 (2.9)	28 (3.4)	7 (1.1)	5 (0.9)
Russian Federation	26 (2.7)	23 (3.4)	29 (3.5)	6 (1.3)	4 (0.8)
Sweden	25 (2.5)	22 (3.2)	27 (3.3)	7 (1.5)	7 (1.2)
Portugal	23 (2.3)	19 (2.9)	28 (3.3)	8 (1.4)	8 (1.2)
Lithuania	22 (2.2)	21 (3.1)	24 (3.0)	5 (0.9)	5 (1.0)
Norway (9)	22 (1.9)	20 (2.6)	24 (3.2)	5 (1.2)	4 (1.0)
England	21 (2.5)	18 (2.9)	25 (3.8)	8 (1.5)	5 (1.0)
Finland	21 (1.7)	19 (2.7)	22 (2.5)	9 (1.1)	6 (1.0)
United Arab Emirates	20 (1.0)	18 (1.7)	22 (1.3)	7 (0.6)	6 (0.4)
Qatar	15 (1.6)	13 (2.0)	17 (3.0)	7 (1.3)	7 (1.3)
Italy	15 (1.7)	13 (2.4)	16 (2.5)	5 (1.0)	5 (1.1)
France	14 (1.8)	12 (2.2)	17 (2.5)	6 (1.3)	10 (1.5)
Chile	14 (2.6)	14 (2.6)	14 (3.2)	5 (1.0)	5 (0.9)
Georgia	13 (1.8)	11 (2.3)	14 (2.5)	5 (1.2)	5 (1.3)
Malaysia	12 (1.0)	12 (1.3)	12 (1.4)	8 (0.9)	9 (1.1)
International Average	**25 (0.5)**	**23 (0.6)**	**28 (0.6)** ▲	**7 (0.2)**	**6 (0.2)**
Benchmarking Participants					
Moscow City, Russian Fed.	38 (2.5)	32 (3.3)	44 (3.4)	5 (0.9)	5 (0.9)
Ontario, Canada	35 (2.4)	34 (4.3)	35 (4.0)	6 (1.2)	7 (1.2)
Dubai, UAE	34 (1.5)	27 (2.4)	41 (2.9)	7 (1.1)	5 (0.9)
Quebec, Canada	26 (2.2)	21 (3.3)	31 (3.5)	8 (1.5)	11 (1.7)
Abu Dhabi, UAE	14 (1.1)	13 (2.0)	15 (1.6)	7 (0.8)	6 (0.5)

▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

For writing the robot's rule, where students were expected to type their answer in the response line, the omit rate jumped to 34 percent. Exhibit 39 presents the percentages of correct responses for 2B. Perhaps due to the high percent of omissions, the average percentage of correct responses—19 percent—was even lower than the average percentage completing the table. Korea (39%), Singapore (35%), and Chinese Taipei (34%) had the best performance. Boys had higher percentages of correct responses than girls on average across countries.

Robots Screen 2B – Percent Correct Overall and by Gender

Country	Percent Correct (3*x* – 1)		
	Overall Country	Girls	Boys
Korea, Rep. of	39 (2.4)	32 (2.9)	44 (3.3)
Singapore	35 (2.0)	29 (2.6)	40 (2.8)
Chinese Taipei	34 (2.1)	28 (2.4)	40 (3.1)
Hong Kong SAR	28 (2.5)	22 (4.4)	32 (3.6)
Hungary	27 (2.1)	25 (3.0)	30 (3.0)
Turkey	27 (2.6)	26 (3.4)	27 (3.5)
Russian Federation	23 (2.6)	23 (4.4)	23 (3.0)
Israel	22 (2.3)	19 (3.2)	24 (3.3)
England	22 (2.9)	19 (3.3)	25 (4.6)
United States	19 (1.8)	16 (2.1)	22 (2.4)
United Arab Emirates	17 (1.1)	15 (1.7)	18 (1.5)
Lithuania	16 (2.1)	17 (2.7)	16 (2.6)
Portugal	16 (1.8)	14 (2.6)	17 (2.4)
Sweden	14 (2.0)	13 (2.6)	14 (3.0)
Norway (9)	13 (1.9)	12 (2.3)	15 (2.6)
Finland	12 (1.4)	12 (1.8)	11 (1.8)
Qatar	11 (1.6)	10 (1.9)	11 (2.4)
Italy	10 (1.4)	10 (1.8)	9 (2.0)
Georgia	8 (1.4)	7 (1.8)	10 (2.2)
France	8 (1.4)	5 (1.5)	12 (2.3)
Chile	4 (0.8)	4 (1.2)	5 (1.1)
Malaysia	3 (0.5)	3 (0.6)	3 (0.9)
International Average	**19 (0.4)**	**17 (0.6)**	**20 (0.6)** ▲
Benchmarking Participants			
Moscow City, Russian Fed.	34 (2.7)	28 (3.6)	38 (3.5)
Dubai, UAE	31 (2.7)	25 (2.8)	36 (4.2)
Ontario, Canada	24 (2.5)	27 (4.5)	22 (3.6)
Quebec, Canada	12 (1.9)	9 (2.1)	15 (3.0)
Abu Dhabi, UAE	9 (1.2)	10 (2.2)	9 (1.3)

Percentage significantly higher than other gender ▲

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Conclusions and Reflections

Building provides a good example of a class of mathematics PSI tasks requiring students to construct a container, structure, or device that will be used for a specific real-world purpose. It also provides a good example of using visuals to help students "see" the path to an end product.

- Tasks based on related step-by-step activities may be susceptible to having a sequence of items all with similar content (e.g., right angle triangle). However, mathematics assessments at upper grade levels involving algebra and geometry are particularly sensitive to variations in curriculum across countries. Requiring items based on a variety of content would provide a greater range of opportunities for students with different curricular backgrounds to succeed.
- The graphical constructed response item requiring students to simulate cutting the building's walls on the board illustrates the promise of digital assessment to go beyond providing an achievement measure. Across the PSI items, the students often were not comfortable when asked to think or reason (e.g., increasing the radius of the rain barrel), but most were engaged enough with cutting the boards to provide process data that can be analyzed to help advance student learning. The student responses to this item also provided grist for research into automated scoring (Appendix C).
- The *Robots* items represent an innovative approach to assessment that was not available in the paper-and-pencil environment. Although these prototypes need further development, the idea can be kept in mind for the future.

CHAPTER 4

Science Grade 8

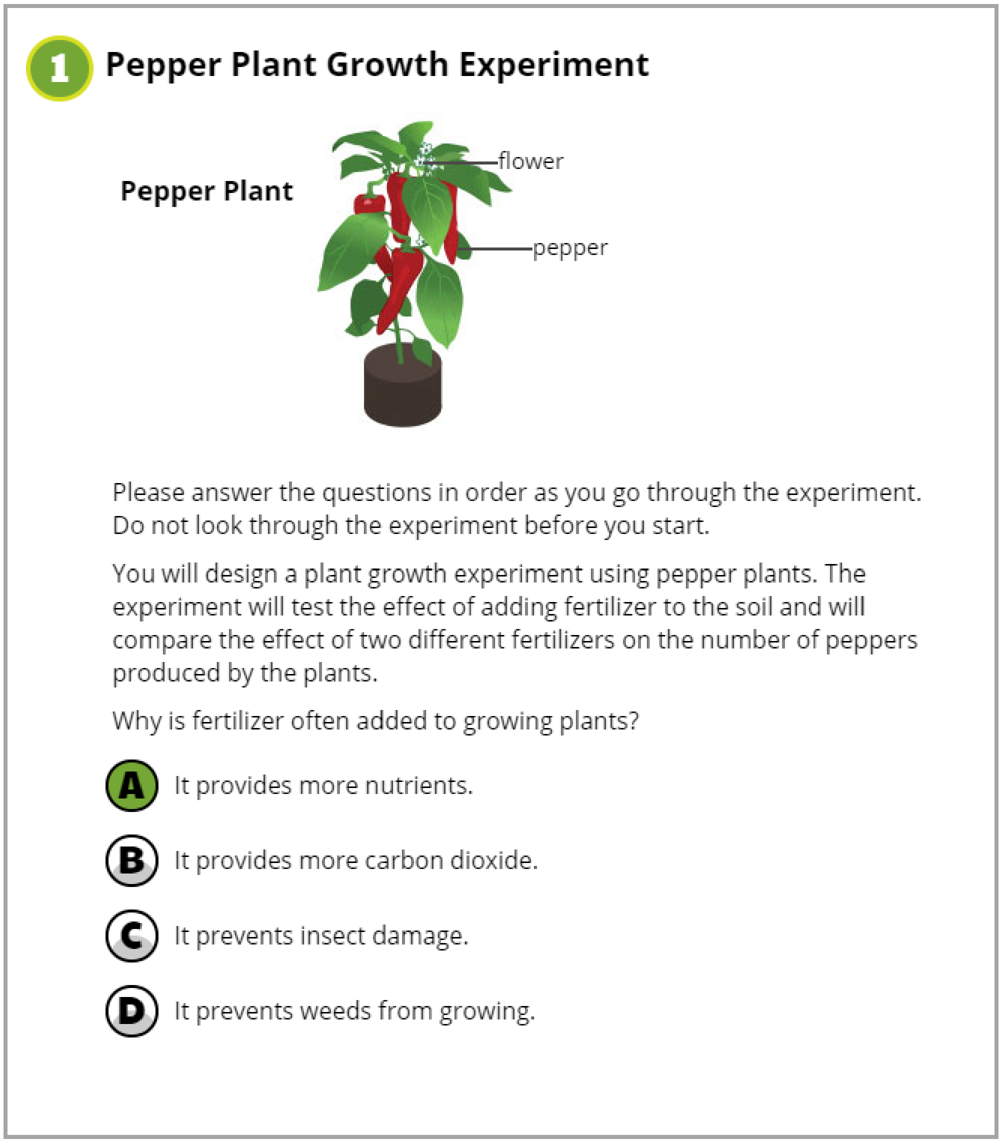
Pepper Plants

About the Task

The *Pepper Plants* task asked eighth grade students to design and analyze the results of an experiment testing the effectiveness of adding fertilizer to the plants' soil and comparing the effect of two different fertilizers on pepper production. Students were asked to create their own experimental setups, which required using their knowledge of basic plant biology combined with their understanding of the principles of experimental design. Innovative response interaction spaces enabled students to control the amounts of the two different fertilizers and water supplied to each of three growth tanks containing the same number of pepper plant seedlings. As they worked through the task, students also answered a variety of multiple-choice and constructed response items. Students were requested to answer the questions in order as they worked through the task, and not to look through the investigation before starting.

Screen 1 – Pepper Plant Growth Experiment

This screen introduces the task of designing a plant growth experiment. It also orients students to the subject of the experiment by asking why fertilizer is added to growing plants. Students obtained credit if they correctly recognized that fertilizers provide nutrients that help plants grow (option A).



1 Pepper Plant Growth Experiment

Pepper Plant

Please answer the questions in order as you go through the experiment. Do not look through the experiment before you start.

You will design a plant growth experiment using pepper plants. The experiment will test the effect of adding fertilizer to the soil and will compare the effect of two different fertilizers on the number of peppers produced by the plants.

Why is fertilizer often added to growing plants?

Ⓐ It provides more nutrients.

Ⓑ It provides more carbon dioxide.

Ⓒ It prevents insect damage.

Ⓓ It prevents weeds from growing.

Maximum Score Points: 1
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Knowing

Exhibit 40 presents the percentage of students, overall and by gender, that chose the correct response option in each country. About three-fourths of the students (78%) answered this item correctly on average across countries. On average, there essentially was no difference overall in performance between boys and girls.

Exhibit 40

Pepper Plants Screen 1 – Percent Correct Overall and by Gender

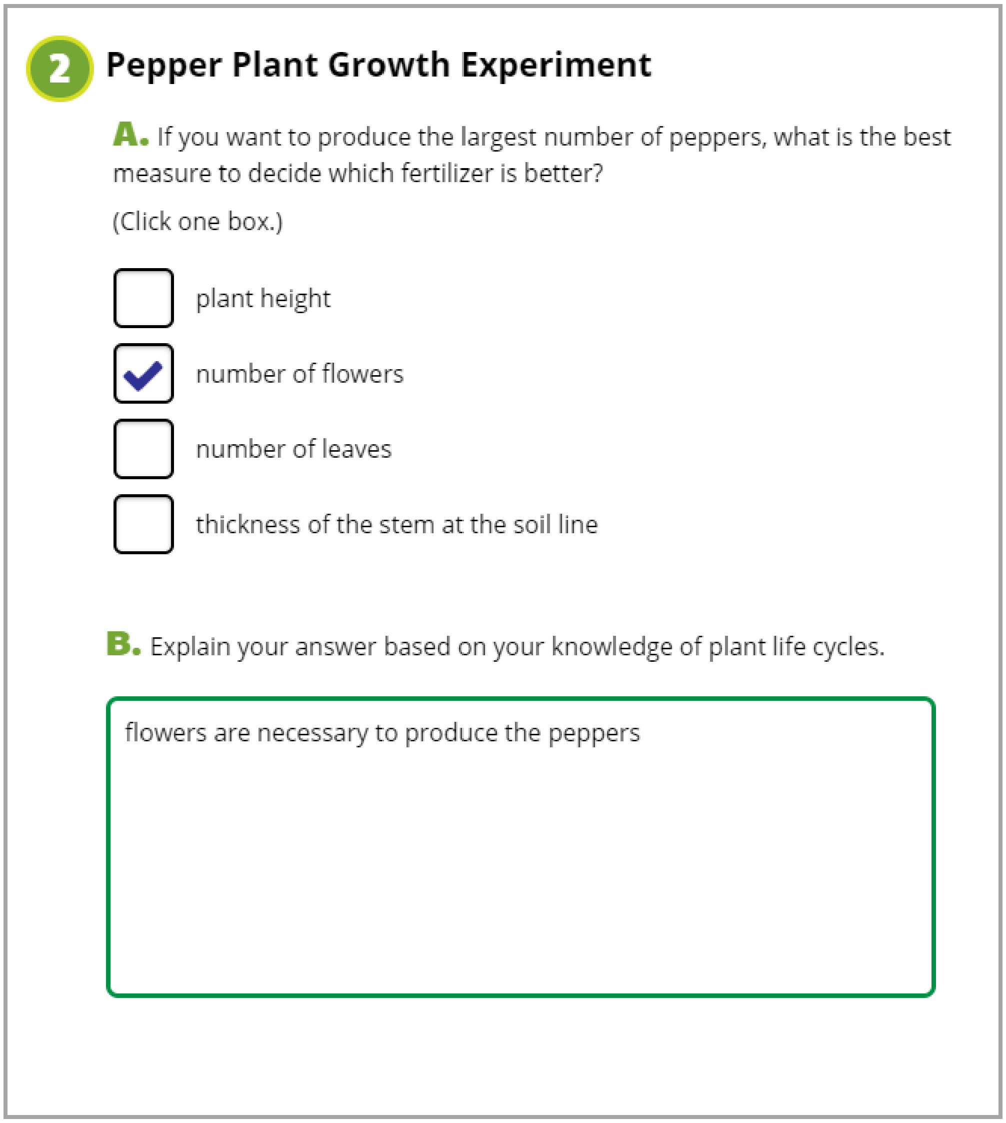
Country	Percent Correct (It provides more nutrients)		
	Overall Country	Girls	Boys
Chinese Taipei	93 (1.2)	92 (1.5)	94 (1.5)
Russian Federation	93 (1.3)	92 (1.4)	93 (2.0)
Singapore	93 (1.2)	93 (1.7)	93 (1.5)
Korea, Rep. of	92 (1.4)	94 (1.5)	90 (2.2)
Finland	90 (1.3)	92 (1.3)	89 (2.0)
United States	85 (1.4)	85 (2.3)	84 (1.5)
Hungary	84 (1.9)	86 (2.5)	82 (2.7)
Norway (9)	83 (1.9)	85 (2.3)	82 (2.4)
Lithuania	80 (1.9)	81 (2.7)	79 (2.8)
England	80 (2.1)	85 (2.5)	75 (3.1)
Israel	79 (2.0)	80 (2.9)	79 (2.6)
Turkey	76 (2.0)	75 (2.9)	76 (2.6)
Malaysia	74 (1.6)	76 (2.5)	73 (2.4)
France	72 (2.2)	73 (2.9)	72 (2.8)
Chile	72 (2.3)	71 (3.5)	73 (3.2)
Sweden	70 (2.6)	73 (3.1)	67 (3.8)
Hong Kong SAR	70 (2.9)	68 (3.9)	71 (3.8)
Italy	67 (2.1)	68 (3.0)	66 (3.1)
Qatar	66 (2.8)	69 (3.5)	62 (3.5)
United Arab Emirates	65 (0.9)	66 (1.3)	65 (1.5)
Georgia	62 (2.8)	65 (3.6)	58 (4.1)
Portugal	61 (2.2)	56 (3.3)	67 (3.4)
International Average	**78 (0.4)**	**78 (0.6)**	**77 (0.6)**
Benchmarking Participants			
Moscow City, Russian Fed.	96 (0.8)	95 (1.5)	97 (0.7)
Quebec, Canada	84 (2.1)	85 (2.8)	82 (3.1)
Ontario, Canada	81 (2.0)	81 (2.7)	80 (2.6)
Dubai, UAE	79 (1.6)	79 (2.1)	79 (2.4)
Abu Dhabi, UAE	54 (1.7)	56 (2.4)	53 (2.5)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 2 – Pepper Plant Growth Experiment

Continuing to encourage students to think about the experiment, this item asks students to identify the best measure of a fertilizer's effectiveness to grow more peppers and to explain their answers based on their knowledge of plant life cycles. Students obtained full credit (2 points) if they selected "number of flowers" and explained that flowers are necessary to produce peppers. Students obtained partial credit (1 point) if they selected "number of flowers" correctly but failed to provide a correct explanation.



2 Pepper Plant Growth Experiment

A. If you want to produce the largest number of peppers, what is the best measure to decide which fertilizer is better?

(Click one box.)

□ plant height
☑ number of flowers
□ number of leaves
□ thickness of the stem at the soil line

B. Explain your answer based on your knowledge of plant life cycles.

flowers are necessary to produce the peppers

Maximum Score Points: 2
Content Domain: Biology
Topic Area: Life Cycles, Reproduction, and Heredity
Cognitive Domain: Applying

Exhibit 41 shows that 18 percent of students on average across countries received full credit by identifying “number of flowers” as the best indicator of number of peppers and providing a correct accompanying explanation. Another 18 percent of students earned partial credit by selecting “number of flowers” without providing a correct explanation. Singapore and Lithuania had the highest average performance with 45 and 31 percent of students earning full credit, respectively. Essentially no overall difference was found between the performance of boys and girls.

Exhibit 41

Pepper Plants Screen 2 – Percent Full Credit Overall and by Gender

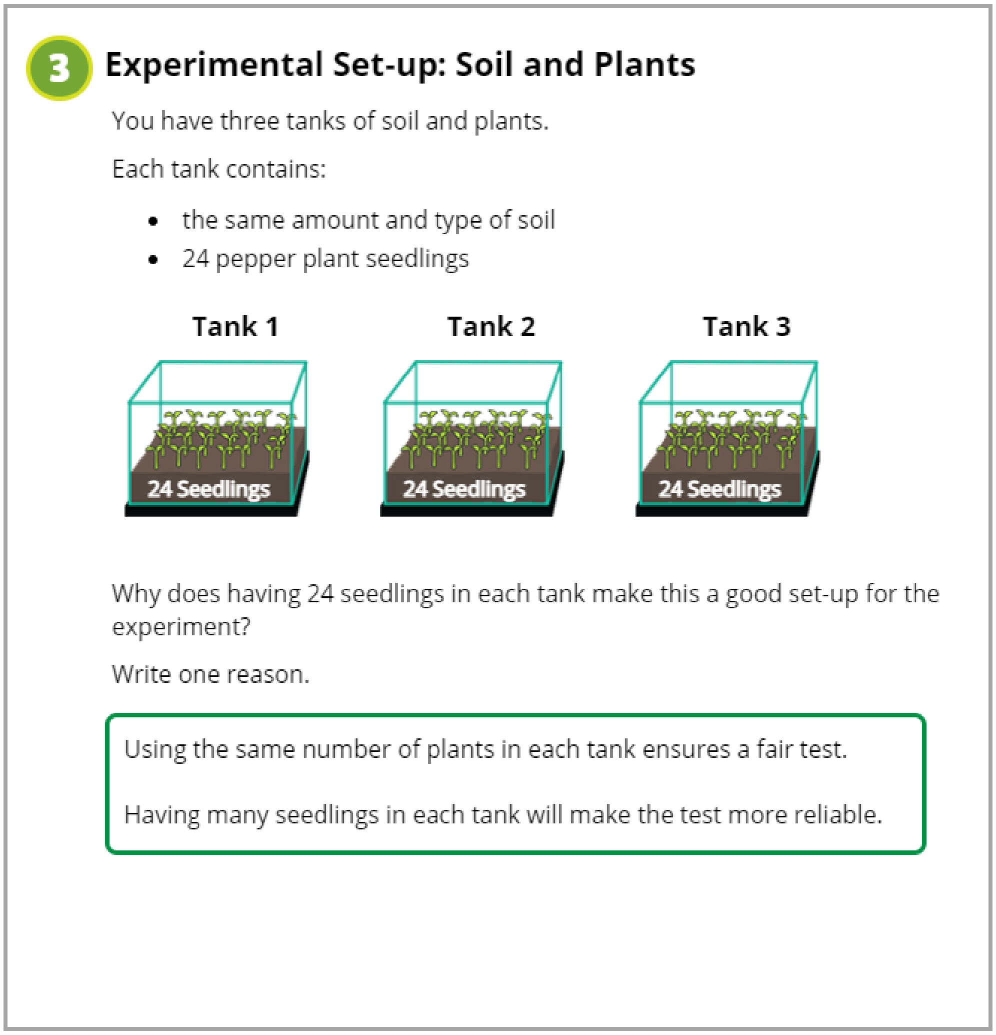
Country	Percent Full Credit (Selects Number of Flowers and Gives Correct Explanation)			Percent Partial Credit (Selects Number of Flowers but No Explanation)
	Overall Country	Girls	Boys	
Singapore	45 (1.9)	42 (2.7)	48 (2.6)	11 (1.2)
Lithuania	31 (2.3)	36 (3.5)	27 (3.8)	26 (2.1)
Chinese Taipei	29 (1.9)	29 (2.5)	30 (2.5)	14 (1.3)
Finland	22 (1.4)	23 (2.2)	22 (2.1)	20 (1.8)
Israel	22 (2.1)	22 (3.1)	22 (2.8)	15 (1.4)
Korea, Rep. of	20 (2.1)	21 (2.6)	20 (2.7)	14 (1.4)
Russian Federation	19 (1.9)	19 (2.3)	20 (2.5)	32 (2.1)
United States	19 (1.3)	20 (1.8)	17 (1.7)	10 (0.9)
Italy	19 (1.9)	17 (2.5)	20 (2.7)	23 (1.8)
Hong Kong SAR	17 (2.0)	16 (2.8)	18 (2.6)	17 (1.7)
Hungary	15 (1.4)	12 (1.8)	18 (2.1)	18 (1.9)
Turkey	14 (1.7)	16 (2.4)	13 (2.1)	17 (1.6)
France	14 (1.7)	14 (2.1)	14 (2.6)	21 (2.2)
Portugal	14 (1.6)	12 (1.9)	16 (2.6)	11 (1.4)
Sweden	14 (1.7)	15 (2.5)	13 (2.3)	19 (1.7)
United Arab Emirates	13 (0.7)	12 (1.1)	14 (1.0)	15 (0.7)
Qatar	11 (1.5)	9 (1.5)	14 (2.7)	12 (1.5)
Chile	11 (1.2)	11 (2.3)	10 (3.0)	11 (1.2)
Malaysia	11 (1.0)	12 (1.7)	9 (1.3)	17 (1.2)
Georgia	10 (1.6)	10 (2.1)	10 (2.2)	18 (2.0)
Norway (9)	8 (1.3)	9 (2.0)	7 (1.6)	35 (2.6)
England	7 (1.1)	8 (1.7)	6 (1.6)	17 (2.1)
International Average	**18 (0.4)**	**17 (0.5)**	**18 (0.5)**	**18 (0.4)**
Benchmarking Participants				
Moscow City, Russian Fed.	32 (2.3)	29 (2.7)	35 (3.1)	17 (1.6)
Quebec, Canada	25 (2.1)	26 (3.0)	24 (3.1)	14 (1.5)
Dubai, UAE	21 (1.7)	19 (2.1)	22 (2.4)	13 (1.5)
Ontario, Canada	13 (1.3)	18 (2.4)	8 (1.8)	11 (1.8)
Abu Dhabi, UAE	8 (1.0)	8 (1.5)	8 (1.4)	17 (1.2)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 3 – Experimental Setup: Soil and Plants

This screen introduces the experimental setup for testing the fertilizers, which consisted of three growth tanks each containing the same amount and type of soil and 24 seedlings. Students were asked to explain why having 24 seedlings in each tank constitutes a good setup for an experiment. Students obtained credit (1 point) either by explaining that having the same number of plants in each tank constitutes a fair test or that having a large number of plants in each tank accounts for plant-to-plant variations (reliability).



3 Experimental Set-up: Soil and Plants

You have three tanks of soil and plants.

Each tank contains:

- the same amount and type of soil
- 24 pepper plant seedlings

Why does having 24 seedlings in each tank make this a good set-up for the experiment?

Write one reason.

Using the same number of plants in each tank ensures a fair test.

Having many seedlings in each tank will make the test more reliable.

Maximum Score Points: 1
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Reasoning
Science Practice: Generating Evidence

Exhibit 42 shows that on average across countries, 27 percent of students answered the item correctly. Explanations based on the nature of a fair test were more common (18%) than explanations based on the reliability of the setup (8%). That almost three-fourths of the eighth grade students did not provide a correct answer was worrisome, but as the task progressed somewhat greater percentages of students displayed understanding of the nature of a fair test. On average, a higher percentage of girls than boys answered this item correctly (29% vs. 24%).

Pepper Plants Screen 3 – Percent Correct Overall and by Gender

Country	Percent Correct (Gives 1 Correct Reason)			Percent by Correct Response Type	
	Overall Country	Girls	Boys	Explanation Refers to Supporting a Fair Test	Explanation Refers to Supporting a Reliable Test
Singapore	69 (1.9)	70 (2.8)	68 (2.6)	54 (1.8)	16 (1.5)
England	54 (2.5)	56 (3.5)	51 (3.0)	48 (2.7)	6 (1.4)
Hong Kong SAR	47 (3.1)	53 (3.9)	43 (4.0)	41 (3.0)	7 (1.3)
United States	44 (1.9)	48 (2.6)	39 (2.5)	22 (1.7)	21 (1.3)
Israel	42 (1.9)	44 (3.2)	40 (3.0)	30 (2.2)	12 (1.7)
Turkey	33 (2.1)	36 (3.1)	29 (2.7)	28 (2.0)	4 (0.8)
Korea, Rep. of	29 (1.9)	34 (3.0)	25 (2.5)	24 (1.8)	5 (0.9)
Chinese Taipei	28 (1.9)	29 (2.6)	26 (2.9)	23 (1.8)	4 (0.7)
Sweden	25 (2.2)	27 (3.2)	23 (3.0)	14 (1.6)	11 (1.5)
Qatar	23 (2.0)	26 (3.0)	21 (2.7)	10 (1.6)	13 (1.6)
Finland	23 (1.5)	28 (2.3)	18 (2.1)	11 (1.1)	11 (1.1)
Hungary	21 (1.8)	23 (2.4)	20 (2.6)	18 (1.7)	4 (0.7)
France	20 (1.8)	20 (2.6)	21 (2.6)	11 (1.4)	9 (1.4)
Russian Federation	19 (2.2)	21 (2.9)	17 (2.6)	11 (1.7)	8 (1.2)
Georgia	19 (2.3)	21 (3.2)	17 (2.8)	4 (1.1)	14 (2.1)
Lithuania	19 (2.1)	21 (3.0)	16 (2.8)	11 (1.7)	7 (1.2)
Malaysia	16 (1.3)	20 (1.7)	13 (1.6)	14 (1.3)	2 (0.5)
United Arab Emirates	15 (0.7)	17 (1.2)	13 (0.8)	11 (0.5)	5 (0.4)
Portugal	13 (1.3)	14 (2.4)	11 (2.1)	9 (1.2)	4 (0.8)
Chile	12 (1.5)	14 (2.3)	9 (1.4)	3 (0.7)	8 (1.5)
Norway (9)	10 (1.6)	12 (2.5)	9 (2.4)	1 (0.4)	9 (1.6)
Italy	9 (1.4)	11 (2.0)	7 (1.8)	4 (1.1)	4 (0.9)
International Average	**27 (0.4)**	**29 (0.6)** ▲	**24 (0.5)**	**18 (0.4)**	**8 (0.3)**
Benchmarking Participants					
Quebec, Canada	33 (2.6)	35 (3.5)	31 (3.5)	11 (1.7)	22 (2.1)
Dubai, UAE	31 (1.6)	32 (2.6)	30 (2.5)	23 (1.4)	8 (1.0)
Ontario, Canada	21 (2.0)	27 (2.9)	16 (2.8)	7 (1.5)	14 (2.7)
Moscow City, Russian Fed.	12 (1.3)	12 (2.1)	13 (1.7)	10 (1.2)	2 (0.6)
Abu Dhabi, UAE	11 (1.0)	14 (1.8)	8 (1.1)	7 (0.9)	4 (0.6)

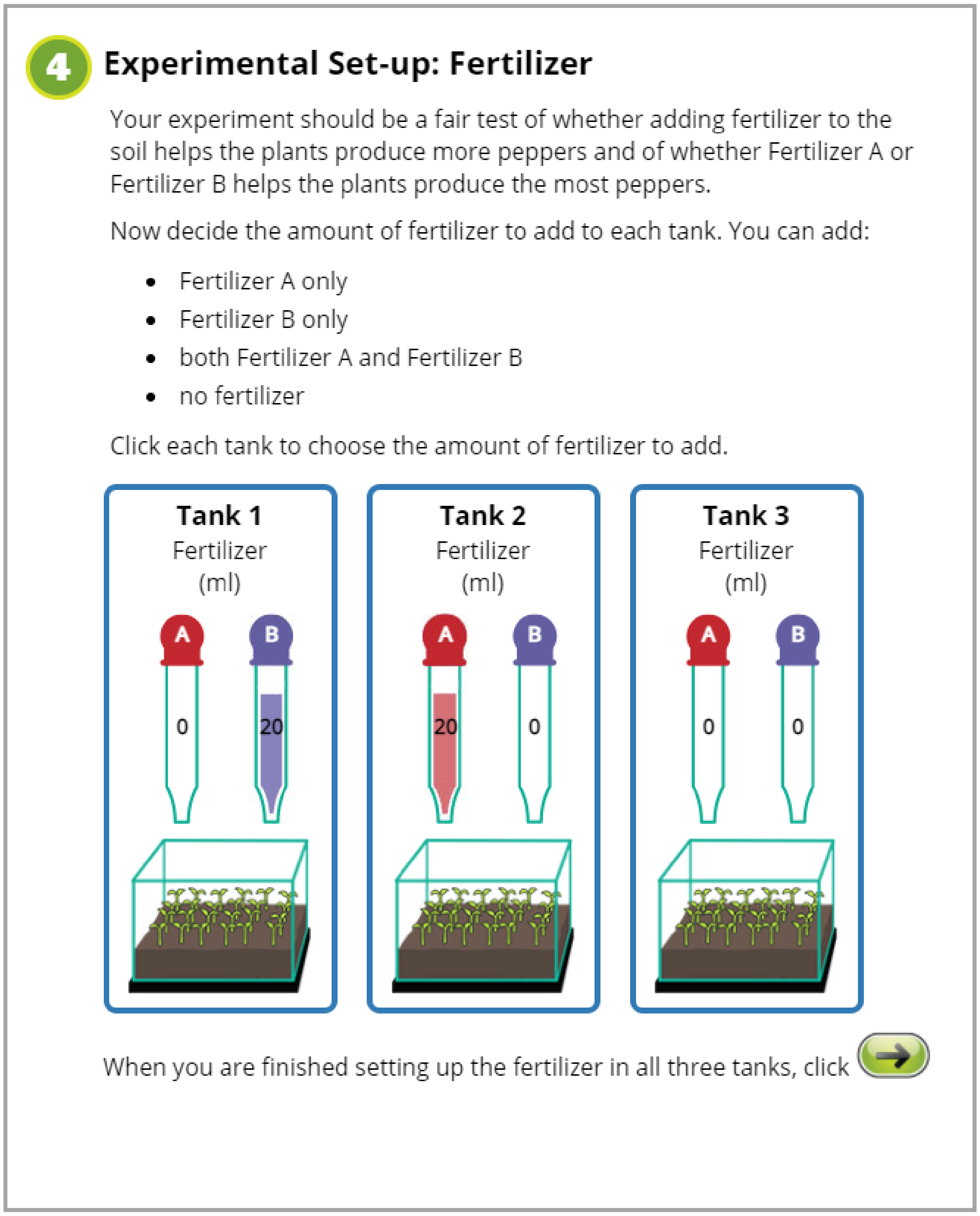
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 4 – Experimental Setup: Fertilizer

On Screen 4, students implemented their experimental design by deciding how much of each of the two fertilizers to add to each tank—20 ml, 10 ml, or 0 ml. To set up the fertilizer combination for a tank, students selected (clicked) the tank and were directed to a screen specific to that tank. In this new screen, students chose the amount of each fertilizer they wished to use in that particular tank.



4 **Experimental Set-up: Fertilizer**

Your experiment should be a fair test of whether adding fertilizer to the soil helps the plants produce more peppers and of whether Fertilizer A or Fertilizer B helps the plants produce the most peppers.

Now decide the amount of fertilizer to add to each tank. You can add:

- Fertilizer A only
- Fertilizer B only
- both Fertilizer A and Fertilizer B
- no fertilizer

Click each tank to choose the amount of fertilizer to add.

When you are finished setting up the fertilizer in all three tanks, click ➔

Maximum Score Points: 2
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Reasoning
Science Practice: Generating Evidence

Students obtained full credit (2 points) for this activity if two conditions were fulfilled:

1) Only Fertilizer A was added to one tank, the same amount of only Fertilizer B was added to a second tank, and

2) No fertilizer was added to the third tank.

In other words, full credit was awarded to responses that showed comprehension of a fair test (adding the same amount of each fertilizer separately to its own tank) and the correct use of a control tank (no fertilizer in a third tank). Students obtained partial credit (1 point) if their response showed comprehension of a fair test but not the appropriate use of a control tank.

Exhibit 43 contains the results for students' success with their experimental setups. Across countries, 43 percent of students, on average, applied the fertilizer correctly by adding the same amount of each of Fertilizer A and of Fertilizer B to separate tanks (e.g., Tanks 1 and 2). However, only 15 percent managed the correct application of the fertilizers **and** also did not add any fertilizer to the third tank so it could be used as a control. A higher percentage of girls than boys, on average across countries, managed all aspects of the experiment's setup correctly (16% vs. 13%).

Exhibit 43

Pepper Plants Screen 4 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Fair Test and Control)			Percent Partial Credit (Fair Test but No Control)
	Overall Country	Girls	Boys	
Singapore	55 (2.0)	60 (2.9)	51 (2.5)	16 (1.4)
Hong Kong SAR	23 (2.4)	25 (3.2)	22 (3.5)	28 (2.3)
Chinese Taipei	22 (1.9)	24 (2.6)	20 (2.6)	32 (2.1)
United States	19 (1.5)	19 (2.1)	20 (2.0)	29 (1.7)
Finland	19 (1.5)	25 (2.3)	13 (1.9)	33 (1.7)
Korea, Rep. of	19 (1.5)	23 (2.5)	15 (2.0)	39 (2.0)
Sweden	18 (2.2)	20 (2.8)	16 (2.8)	28 (2.4)
England	18 (2.1)	21 (2.9)	15 (2.5)	28 (2.0)
Turkey	16 (1.6)	17 (2.3)	16 (2.4)	20 (1.8)
Russian Federation	16 (1.7)	17 (2.3)	15 (2.2)	38 (2.1)
Norway (9)	15 (2.0)	16 (2.6)	15 (2.6)	29 (2.0)
Lithuania	15 (1.6)	19 (2.4)	10 (2.1)	34 (2.1)
Israel	14 (1.9)	14 (2.7)	13 (2.4)	24 (1.8)
United Arab Emirates	10 (0.6)	10 (0.7)	10 (1.0)	19 (0.8)
France	9 (1.4)	11 (2.1)	8 (1.7)	34 (2.1)
Hungary	8 (0.9)	7 (1.5)	9 (1.6)	35 (2.3)
Portugal	8 (1.3)	6 (1.7)	9 (2.2)	28 (2.2)
Qatar	6 (1.3)	7 (1.5)	6 (2.2)	21 (2.2)
Italy	5 (0.9)	5 (1.3)	5 (1.6)	33 (2.2)
Chile	5 (0.8)	5 (1.2)	4 (1.2)	27 (2.0)
Malaysia	4 (0.6)	4 (0.9)	4 (0.7)	22 (1.6)
Georgia	1 (0.6)	1 (0.9)	1 (0.5)	20 (2.1)
International Average	**15 (0.3)**	**16 (0.5)** ▲	**13 (0.5)**	**28 (0.4)**
Benchmarking Participants				
Moscow City, Russian Fed.	27 (2.2)	22 (3.1)	31 (3.0)	34 (2.0)
Dubai, UAE	23 (1.4)	22 (2.1)	24 (2.2)	25 (2.0)
Ontario, Canada	17 (2.2)	21 (3.0)	14 (2.9)	36 (3.0)
Quebec, Canada	14 (1.6)	17 (2.5)	12 (1.9)	42 (2.9)
Abu Dhabi, UAE	5 (0.6)	5 (1.0)	5 (0.9)	16 (1.2)

▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

The combined percentages of full and partial credit provided in Exhibit 43, reveal that more than two-fifths of students (43%), on average across countries, provided fertilizer setups that demonstrated understanding of a fair test. However, there still were relatively large percentages of students in a number of countries that did not demonstrate understanding of a fair test. To provide further information about what types of responses these students did provide, the process data were investigated for evidence of misunderstanding the task or misconceptions about what constituted a fair test in this investigation.

Exhibit 44 illustrates the two correct setups for demonstrating understanding of a fair test to determine the most effective fertilizer (A or B). That is, students added each fertilizer to a different tank keeping the amount constant across the two tanks.

Exhibit 44 also illustrates three incorrect setups that might reflect misunderstandings about the goal of the experiment, misconceptions about a fair test, or general confusion:

1) tests if more fertilizer is better (students adding the same amount of both fertilizers to one tank and twice as much of both fertilizers to a second tank),

2) adds an equal amount of both fertilizers to two tanks (applying the same amount of both fertilizers to both tanks may have seemed fair, but it resulted in no variation in type or amount of fertilizer), or

3) varies both the type and the amount (adding an amount of one fertilizer to one tank and a different amount of the other fertilizer to a second tank, making comparisons impossible).

Exhibit 44: *Pepper Plants* Screen 4 – Understanding of a Fair Test

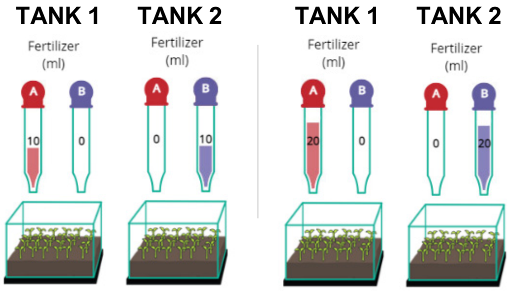
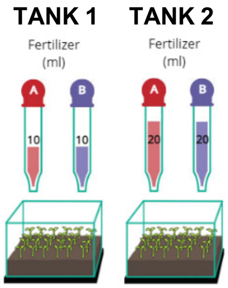
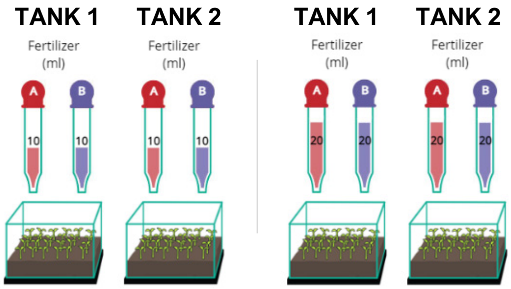
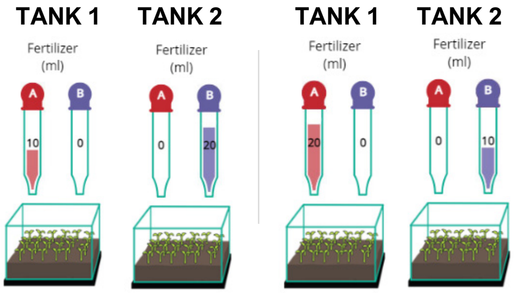
Student Setup	Examples
Fair Test Student understands the nature of a fair test, adds equal amounts of one fertilizer at a time to separate tanks.	 TANK 1 TANK 2 TANK 1 TANK 2 Fertilizer (ml) A 10 B 0 A 0 B 10 A 20 B 0 A 0 B 20
Incorrect Setups	
(1) Varies Total Amount Student adds equal amounts of both fertilizers to each tank, but twice as much fertilizer to the second tank.	 TANK 1 TANK 2 Fertilizer (ml) A 10 B 10 A 20 B 20
(2) Adds Equal Total Amount Student adds the same amount of both fertilizers to both tanks so may seem fair, but no variation to provide comparison.	 TANK 1 TANK 2 TANK 1 TANK 2 Fertilizer (ml) A 10 B 10 A 10 B 10 A 20 B 20 A 20 B 20
(3) Varies Type and Amount Student adds one fertilizer at a time to separate tanks but adds them in different amounts so comparison is not possible.	 TANK 1 TANK 2 TANK 1 TANK 2 Fertilizer (ml) A 10 B 0 A 0 B 20 A 20 B 0 A 0 B 10

Exhibit 45 shows that 43 percent of students, on average across countries, demonstrated understanding of a fair test, correctly adding only one fertilizer per tank using the same amount of fertilizer in both tanks. The most common incorrect setup (14% on average) disregarded the difference between fertilizers, applying both fertilizers in one tank and twice as much of both in a second tank. These students may have misread or misunderstood the task and thought it only was about determining whether adding fertilizer helps pepper production, which was the first purpose of the experiment described in the task. The next most common incorrect setup (7% on average) involved adding the same amount of both fertilizers to each of two tanks, possibly indicating a misconception about a fair test. The incorrect setup where students had nothing to compare because they varied the fertilizer and the amount was relatively rare (2% on average).

Exhibit 45

Pepper Plants Screen 4 – Percent of Students with Correct and Incorrect Understanding of a Fair Test

Country	Percent Correct Fair Test (Full and Partial Credit)	Percent Incorrect Setups Indicating Misunderstandings or Misconceptions		
		Varies Total Amount	Adds Equal Total Amount	Varies Type and Amount
Singapore	72 (1.9)	10 (1.0)	3 (0.7)	1 (0.3)
Korea, Rep. of	58 (2.1)	12 (1.5)	8 (1.1)	1 (0.4)
Chinese Taipei	54 (2.1)	15 (1.1)	5 (0.8)	0 (0.2)
Russian Federation	54 (2.4)	11 (1.3)	5 (0.9)	1 (0.5)
Finland	52 (1.9)	13 (1.3)	5 (0.8)	4 (0.7)
Hong Kong SAR	52 (2.4)	15 (1.8)	3 (0.9)	0 (0.0)
Lithuania	48 (2.6)	17 (2.0)	5 (1.0)	2 (0.6)
United States	48 (1.9)	20 (1.3)	4 (0.6)	2 (0.4)
Sweden	46 (3.1)	11 (1.5)	7 (1.0)	2 (0.6)
England	46 (2.6)	17 (1.8)	11 (1.4)	2 (0.6)
Norway (9)	44 (2.6)	12 (1.4)	4 (1.0)	4 (1.0)
France	43 (2.3)	16 (1.7)	9 (1.3)	2 (0.7)
Hungary	43 (2.3)	9 (1.2)	4 (0.8)	3 (0.8)
Israel	38 (2.1)	16 (1.7)	4 (1.1)	2 (0.6)
Italy	38 (2.4)	10 (1.2)	11 (1.3)	3 (0.8)
Turkey	37 (2.4)	11 (1.2)	5 (1.1)	2 (0.6)
Portugal	36 (2.5)	15 (1.7)	7 (1.3)	2 (0.6)
Chile	31 (2.0)	16 (1.7)	9 (1.4)	4 (0.8)
United Arab Emirates	29 (0.9)	13 (0.8)	9 (0.7)	2 (0.2)
Qatar	27 (2.3)	14 (1.6)	11 (1.5)	1 (0.4)
Malaysia	26 (1.6)	22 (1.5)	6 (0.9)	3 (0.6)
Georgia	21 (2.1)	11 (1.8)	13 (1.9)	2 (0.7)
International Average	**43 (0.5)**	**14 (0.3)**	**7 (0.2)**	**2 (0.1)**
Benchmarking Participants				
Moscow City, Russian Fed.	61 (2.3)	11 (1.5)	3 (0.8)	2 (0.7)
Quebec, Canada	56 (2.8)	11 (1.7)	5 (1.1)	2 (0.8)
Ontario, Canada	53 (2.7)	14 (1.8)	5 (0.9)	2 (0.7)
Dubai, UAE	48 (1.8)	13 (1.7)	6 (0.8)	1 (0.4)
Abu Dhabi, UAE	21 (1.2)	15 (1.3)	8 (0.9)	2 (0.5)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Exhibit 46 shows the percentage of students who included 1) a correct control tank (0 ml of both fertilizers), 2) an incorrect control tank with an equal amount of both fertilizers, showing a possible misconception, or 3) correctly used two tanks for the experiment but did not use the third tank. Except in Singapore (65%), correct use of a control tank was not very common, ranging from 7 to 37 percent across the rest of the countries. Just 25 percent of students, on average, correctly used the third tank by not adding any fertilizer to it. Interestingly, 34 percent of students on average across countries provided evidence of a misconception about a control by adding an equal amount of each of the fertilizers to their third tank.

Pepper Plants Screen 4 – Percent of Students with Correct and Incorrect Understanding of a Control

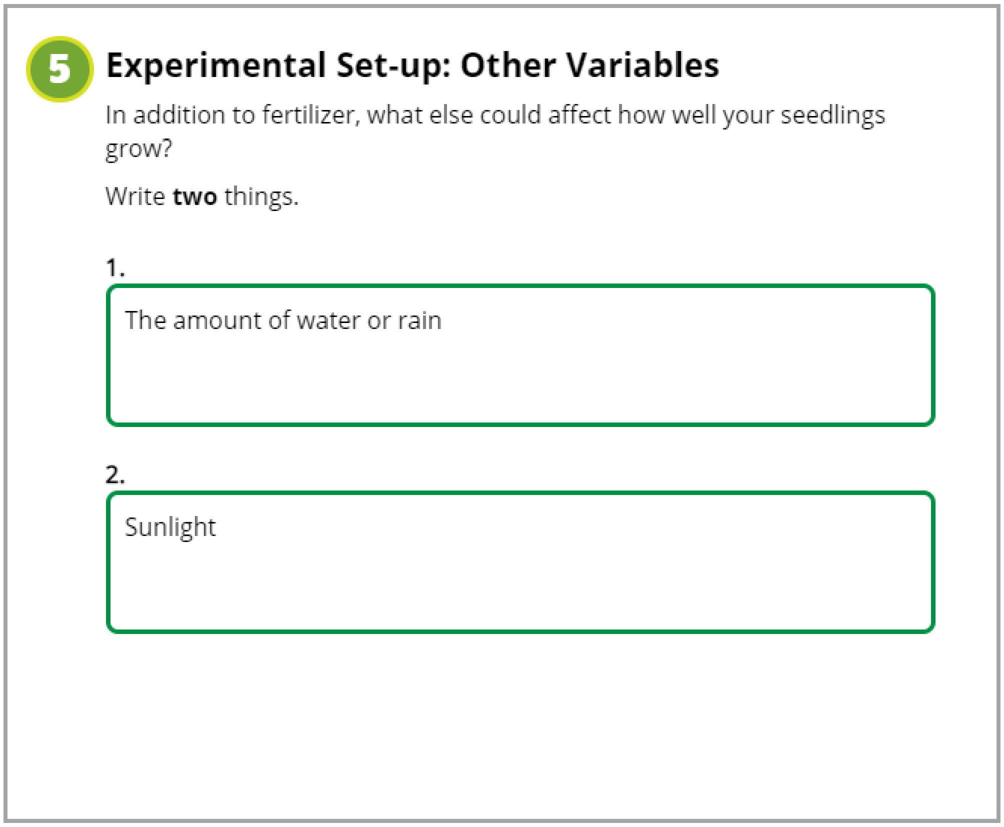
Country	Correct Control Tank (No Fertilizer)	Incorrect Control (Equal Amount Both Fertilizers)	Understands Fair Test but Ignores Third Tank
Singapore	65 (1.9)	19 (1.6)	2 (0.6)
Hong Kong SAR	37 (2.6)	30 (2.4)	3 (1.1)
Chinese Taipei	34 (2.2)	37 (2.2)	4 (0.7)
United States	34 (1.6)	31 (1.8)	9 (1.0)
England	32 (2.3)	39 (2.2)	5 (1.2)
Korea, Rep. of	30 (1.8)	47 (2.0)	4 (0.7)
Finland	29 (1.7)	39 (1.8)	7 (1.1)
Lithuania	27 (2.1)	39 (2.0)	8 (1.2)
Israel	26 (2.5)	27 (2.2)	7 (1.2)
Sweden	25 (2.4)	35 (2.4)	6 (1.0)
Norway (9)	24 (2.3)	33 (2.1)	7 (1.2)
Russian Federation	24 (1.7)	43 (2.1)	5 (1.3)
Turkey	22 (1.8)	23 (1.8)	10 (1.1)
France	22 (2.0)	43 (2.1)	6 (1.1)
Malaysia	21 (1.6)	26 (1.7)	10 (1.1)
United Arab Emirates	19 (0.9)	28 (0.8)	6 (0.5)
Portugal	18 (2.2)	35 (2.3)	7 (1.2)
Qatar	17 (1.9)	31 (2.1)	5 (1.2)
Chile	14 (1.5)	36 (2.2)	10 (1.3)
Hungary	13 (1.2)	39 (2.3)	7 (1.3)
Italy	10 (1.4)	41 (2.4)	11 (1.4)
Georgia	7 (1.5)	31 (2.8)	10 (1.7)
International Average	**25 (0.4)**	**34 (0.4)**	**7 (0.2)**
Benchmarking Participants			
Moscow City, Russian Fed.	36 (2.2)	37 (2.0)	5 (1.0)
Dubai, UAE	33 (2.2)	31 (1.9)	4 (0.8)
Ontario, Canada	29 (2.9)	39 (3.0)	5 (0.9)
Quebec, Canada	23 (2.1)	47 (2.8)	5 (1.1)
Abu Dhabi, UAE	15 (1.4)	23 (1.2)	9 (0.9)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 5 – Experimental Setup: Other Variables

Following Screen 4 which dealt with the central experimental design (assigning amounts of the two fertilizers to experimental and control tanks), Screen 5 asked students to consider other factors that could affect the growth of their seedlings. Students were awarded full credit (2 points) for providing two correct responses (e.g., amount of light, enough water, temperature) and partial credit (1 point) if only one correct response was given.



5 **Experimental Set-up: Other Variables**

In addition to fertilizer, what else could affect how well your seedlings grow?

Write **two** things.

1\.

The amount of water or rain

2\.

Sunlight

Maximum Score Points: 2
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Knowing

Exhibit 47 shows that across countries on average, 54 percent of students correctly provided two things other than fertilizer that could affect the growth of their pepper plants, and a further 21 percent provided one thing. In Finland, the United States, and Singapore, 70 percent of students or more obtained full credit in this item. Girls performed better than boys, on average across countries (58% vs. 51%).

Pepper Plants Screen 5 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Gives 2 Correct Things)			Percent Partial Credit (Gives Only 1 Correct Thing)
	Overall Country	Girls	Boys	
Finland	75 (1.6)	79 (2.4)	72 (2.2)	15 (1.2)
United States	74 (1.2)	76 (1.8)	72 (2.0)	10 (0.9)
Singapore	70 (1.6)	72 (2.2)	69 (2.5)	18 (1.5)
Norway (9)	69 (2.0)	73 (2.8)	65 (3.0)	16 (1.6)
England	68 (2.6)	69 (3.2)	67 (3.4)	14 (1.7)
Turkey	65 (2.0)	69 (2.7)	61 (3.1)	15 (1.5)
Sweden	64 (2.5)	70 (3.5)	59 (4.0)	15 (1.7)
Chinese Taipei	61 (2.3)	64 (3.2)	58 (3.0)	23 (1.8)
Hungary	60 (2.4)	61 (3.4)	58 (3.4)	22 (1.6)
Korea, Rep. of	59 (2.2)	63 (3.0)	56 (3.4)	19 (1.8)
France	55 (2.3)	59 (2.9)	51 (3.5)	19 (1.8)
Chile	55 (2.6)	56 (3.1)	53 (3.8)	21 (1.7)
Israel	53 (2.1)	55 (3.4)	50 (2.8)	15 (1.5)
Hong Kong SAR	51 (2.7)	52 (4.3)	51 (4.1)	18 (1.6)
Russian Federation	47 (2.2)	52 (3.3)	42 (3.3)	30 (2.1)
Lithuania	46 (2.6)	50 (3.5)	41 (3.6)	33 (2.6)
Italy	42 (2.5)	47 (3.5)	37 (3.3)	28 (2.2)
Portugal	39 (2.1)	45 (3.8)	34 (3.2)	29 (2.1)
Malaysia	39 (1.9)	44 (2.2)	34 (2.7)	27 (1.8)
United Arab Emirates	38 (1.0)	44 (1.5)	33 (1.3)	16 (0.7)
Qatar	37 (1.9)	38 (3.0)	35 (3.0)	21 (2.0)
Georgia	31 (2.9)	32 (4.3)	29 (3.4)	35 (2.8)
International Average	**54 (0.5)**	**58 (0.7)** ▲	**51 (0.7)**	**21 (0.4)**
Benchmarking Participants				
Ontario, Canada	71 (2.1)	76 (2.7)	68 (3.2)	13 (1.8)
Quebec, Canada	64 (2.6)	69 (3.1)	59 (4.0)	22 (2.2)
Dubai, UAE	56 (2.0)	58 (3.3)	54 (2.7)	14 (1.4)
Moscow City, Russian Fed.	53 (2.2)	58 (3.7)	49 (2.6)	26 (1.8)
Abu Dhabi, UAE	34 (1.5)	44 (2.3)	25 (1.6)	16 (1.4)

▲ Percentage significantly higher than other gender

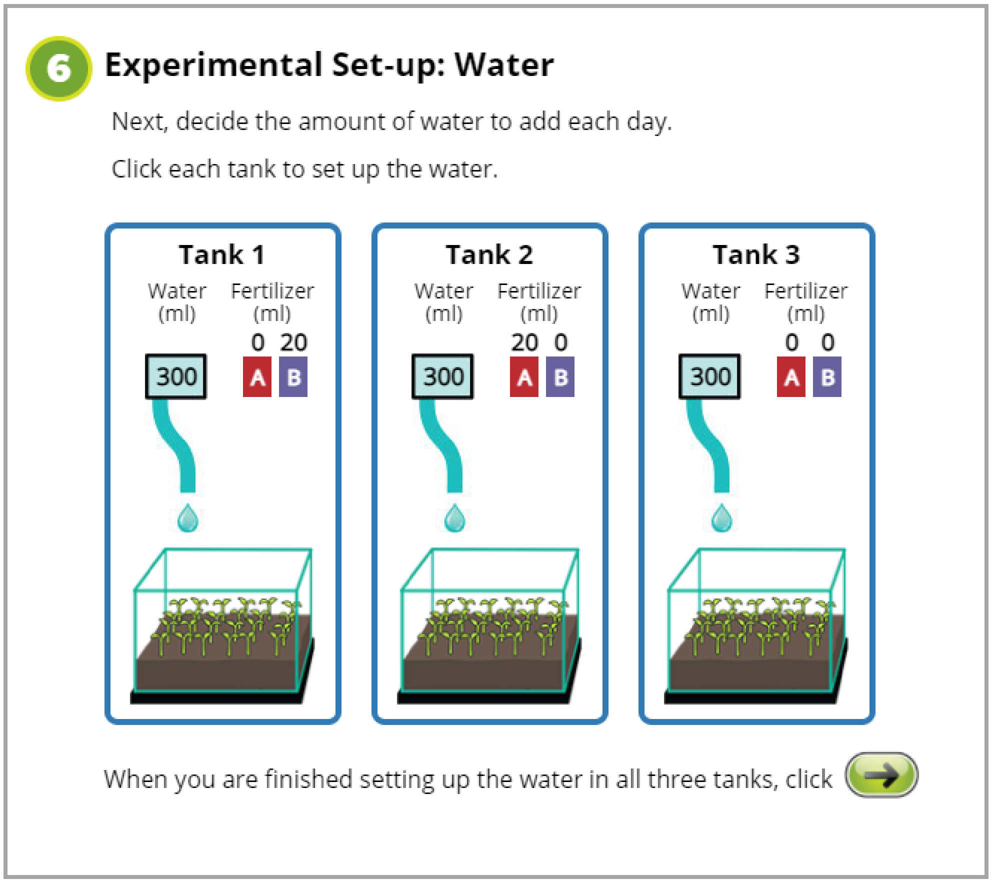
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 6 – Experimental Setup: Water

Screen 6 asked students to decide how much water to add to each of the three tanks. Students could choose to supply 600 ml, 300 ml, or 0 ml of water to each tank. Because the amount of water is a factor that should be held constant across tanks in this experiment, a correct response required setting up the same (non-zero) amount of water for each of the three tanks.

Similar to the procedure for setting up the fertilizer, students selected one growth tank at a time and were referred to a tank-specific screen where they chose the amount of water to supply to that tank.



6 **Experimental Set-up: Water**

Next, decide the amount of water to add each day.

Click each tank to set up the water.

	Tank 1	Tank 2	Tank 3
Water (ml)	300	300	300
Fertilizer (ml) A	0	20	0
Fertilizer (ml) B	20	0	0

When you are finished setting up the water in all three tanks, click

Maximum Score Points: 1
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Reasoning
Science Practice: Generating Evidence

Exhibit 48 shows that 35 percent of students, across countries on average, correctly supplied the same amount of water to all three tanks. On average across countries, girls performed better than boys. Among the incorrect responses, 25 percent of students set two tanks to have the same amount of water.

Exhibit 48

Pepper Plants Screen 6 – Percent Correct Overall and by Gender

Country	Percent Correct (Same Amount of Water for All 3 Tanks)			Percent with Same Amount of Water for Only 2 Tanks (Incorrect)	Percent with Different Amount of Water for All 3 Tanks (Incorrect)
	Overall Country	Girls	Boys		
Singapore	73 (1.9)	77 (2.2)	69 (2.7)	8 (0.9)	12 (1.4)
Chinese Taipei	48 (2.4)	51 (2.7)	45 (3.5)	16 (1.4)	23 (1.9)
Korea, Rep. of	48 (2.2)	50 (3.3)	46 (3.1)	15 (1.5)	25 (1.9)
England	46 (2.8)	48 (3.3)	44 (3.8)	16 (1.8)	25 (2.0)
Hong Kong SAR	44 (2.3)	44 (3.6)	44 (3.4)	9 (1.6)	25 (1.8)
Sweden	41 (2.4)	47 (3.8)	36 (3.5)	25 (2.0)	18 (1.7)
United States	39 (2.2)	39 (2.8)	39 (2.6)	19 (1.6)	30 (2.0)
Norway (9)	37 (2.8)	34 (3.8)	39 (3.8)	24 (2.4)	25 (2.2)
Finland	34 (1.8)	40 (2.8)	29 (2.5)	30 (1.8)	26 (1.7)
Israel	34 (2.2)	35 (3.3)	33 (3.2)	17 (1.8)	24 (1.7)
Turkey	32 (1.6)	33 (2.4)	31 (2.7)	22 (1.7)	22 (2.0)
Russian Federation	32 (1.9)	38 (3.4)	26 (2.3)	36 (2.3)	18 (1.6)
Hungary	31 (2.3)	29 (3.4)	32 (2.8)	34 (2.3)	19 (1.8)
Lithuania	29 (1.9)	32 (3.0)	27 (2.8)	38 (2.0)	23 (2.0)
United Arab Emirates	28 (1.1)	29 (1.4)	26 (1.7)	24 (0.8)	19 (1.0)
Qatar	27 (2.6)	31 (3.8)	24 (3.5)	23 (1.9)	25 (2.5)
Portugal	26 (2.5)	26 (3.4)	25 (3.2)	31 (2.2)	27 (2.0)
France	26 (1.7)	27 (2.6)	24 (2.6)	28 (2.2)	33 (2.1)
Chile	24 (2.0)	22 (2.5)	25 (3.7)	35 (3.3)	25 (2.2)
Italy	23 (1.9)	26 (2.9)	20 (2.6)	39 (2.3)	19 (2.1)
Georgia	21 (1.8)	21 (2.5)	21 (2.7)	31 (2.6)	15 (2.3)
Malaysia	20 (1.6)	21 (2.1)	20 (2.1)	31 (1.8)	34 (1.7)
International Average	**35 (0.5)**	**36 (0.6)** ▲	**33 (0.6)**	**25 (0.4)**	**23 (0.4)**
Benchmarking Participants					
Dubai, UAE	45 (2.7)	45 (3.6)	44 (4.0)	19 (1.6)	19 (2.1)
Moscow City, Russian Fed.	42 (2.4)	40 (3.2)	43 (3.0)	28 (2.3)	18 (1.7)
Ontario, Canada	38 (2.6)	44 (3.4)	33 (3.6)	25 (2.4)	23 (2.2)
Quebec, Canada	38 (2.7)	41 (3.4)	34 (3.7)	30 (2.1)	22 (2.1)
Abu Dhabi, UAE	19 (1.0)	20 (1.7)	17 (1.5)	26 (1.1)	21 (1.4)

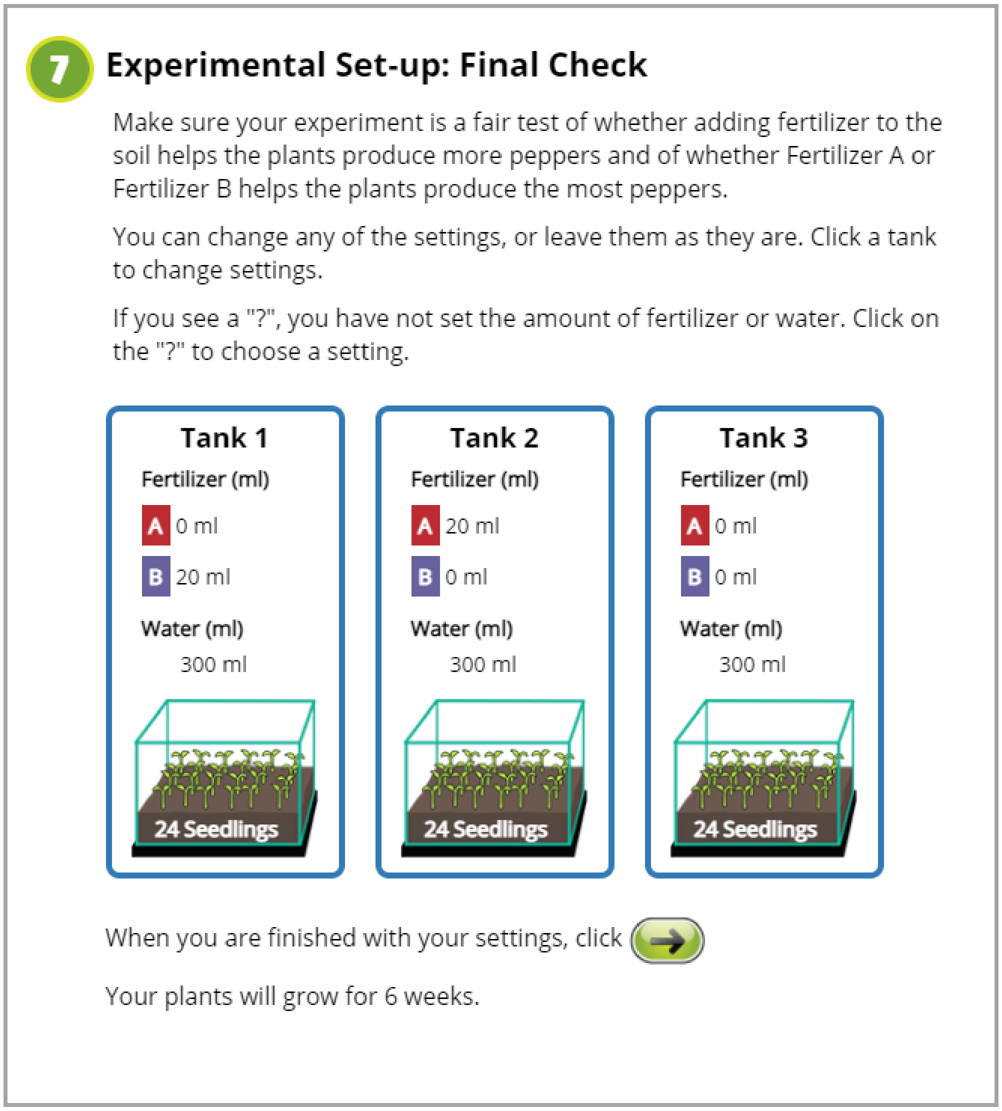
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 7 – Experimental Setup: Final Check

Screen 7 provides students with an opportunity to review the choices they have made in setting up their experiment, and to make any necessary changes. In this screen, students could modify any fertilizer or water choice they wanted. To make a change to a tank's settings, students clicked on the panel for that tank on the screen, and were taken to a tank-specific screen where they could modify both the fertilizer and water settings for that tank.



7 Experimental Set-up: Final Check

Make sure your experiment is a fair test of whether adding fertilizer to the soil helps the plants produce more peppers and of whether Fertilizer A or Fertilizer B helps the plants produce the most peppers.

You can change any of the settings, or leave them as they are. Click a tank to change settings.

If you see a "?", you have not set the amount of fertilizer or water. Click on the "?" to choose a setting.

Tank 1
Fertilizer (ml)
A 0 ml
B 20 ml
Water (ml)
300 ml
24 Seedlings

Tank 2
Fertilizer (ml)
A 20 ml
B 0 ml
Water (ml)
300 ml
24 Seedlings

Tank 3
Fertilizer (ml)
A 0 ml
B 0 ml
Water (ml)
300 ml
24 Seedlings

When you are finished with your settings, click ➔

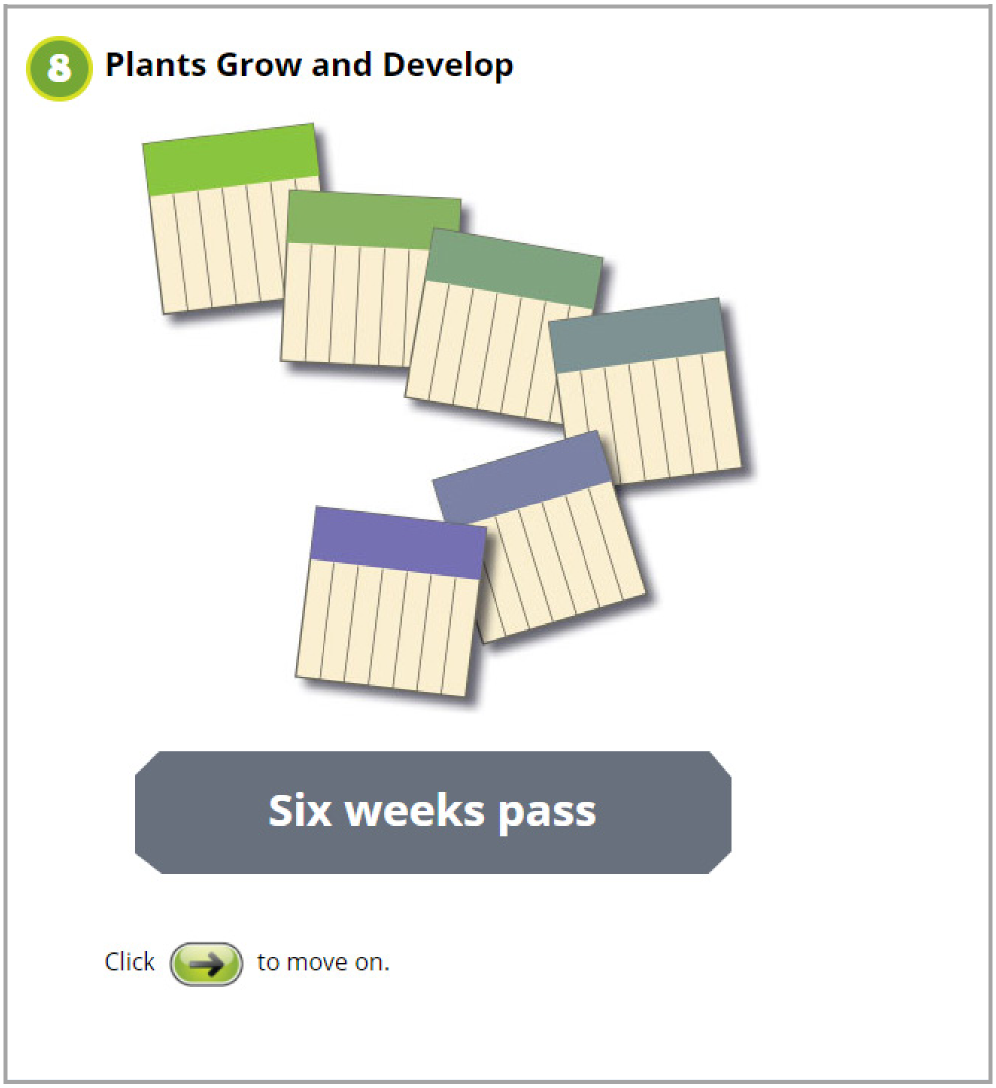
Your plants will grow for 6 weeks.

Maximum Score Points: *Excluded from Scaling and Analysis*
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Reasoning
Science Practice: Generating Evidence

Analyses of the process data revealed that on average across countries 26 percent of students made at least one change (5% changed both fertilizer and water, 14% only changed fertilizer, and 7% only changed water). About two-thirds of the students (67%) selected amounts of fertilizer and water (Screens 4 and 6) and did not make any changes to his or her experimental setup on Screen 7. These students' responses were carried forward from Screens 4 and 6, respectively.

Screen 8 – Plants Grow and Develop

This screen acts as a transition screen indicating to students that six weeks have passed and their plants have grown based on their experimental setups. There are no items on this screen.



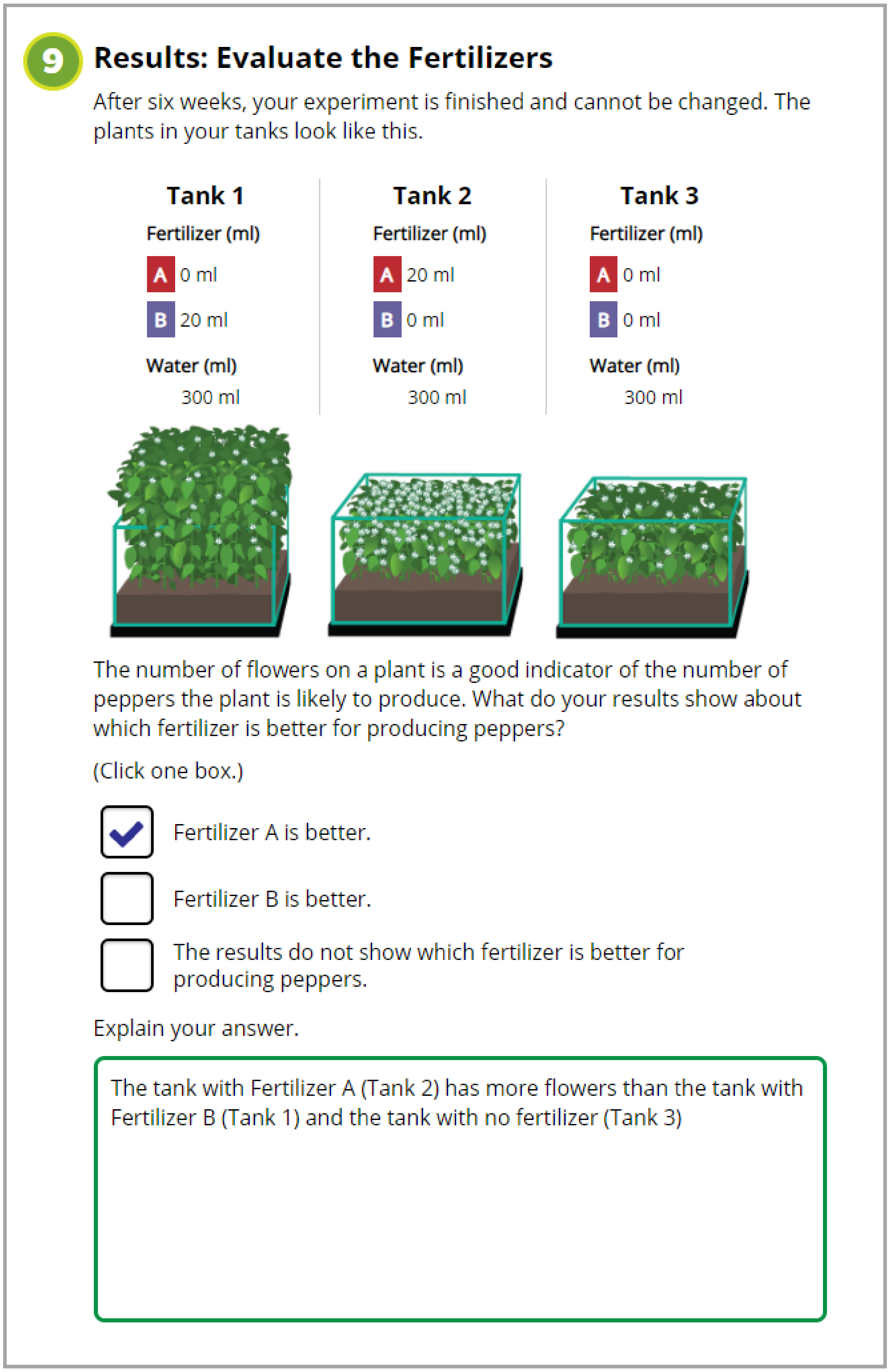
8 Plants Grow and Develop

Six weeks pass

Click → to move on.

Screen 9 – Results: Evaluate the Fertilizers

Screen 9 presented the results of the student's seedling growth in each of the three tanks based on the student's choices of fertilizer and water. The students were reminded that number of flowers on a plant is good indicator of potential pepper production. Then they were asked to interpret their results about the effectiveness of Fertilizer A and Fertilizer B.



9 Results: Evaluate the Fertilizers

After six weeks, your experiment is finished and cannot be changed. The plants in your tanks look like this.

Tank 1	Tank 2	Tank 3
Fertilizer (ml)	Fertilizer (ml)	Fertilizer (ml)
A 0 ml	A 20 ml	A 0 ml
B 20 ml	B 0 ml	B 0 ml
Water (ml)	Water (ml)	Water (ml)
300 ml	300 ml	300 ml

The number of flowers on a plant is a good indicator of the number of peppers the plant is likely to produce. What do your results show about which fertilizer is better for producing peppers?

(Click one box.)

☑ Fertilizer A is better.

□ Fertilizer B is better.

□ The results do not show which fertilizer is better for producing peppers.

Explain your answer.

The tank with Fertilizer A (Tank 2) has more flowers than the tank with Fertilizer B (Tank 1) and the tank with no fertilizer (Tank 3)

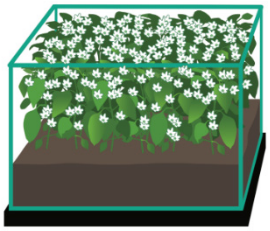
Maximum Score Points: 1
Content Domain: Biology
Topic Area: Life Cycles, Reproduction, and Heredity
Cognitive Domain: Reasoning
Science Practice: Answering the Research Question

Each student's three growth tanks showed plants of various heights and number of flowers based on the student's own particular fertilizer set up. The plants given either 300 or 600 ml of water had the same growth, but those with 0 water withered and died. When students saw their results on Screen 9, they were able to change their experimental setup in an effort to improve their results. Interestingly, analyses of the process data indicated that after seeing Screen 9, on average across countries, 29 percent of the students went back and made changes to both their fertilizer and water settings and another 1 percent adjusted their fertilizer settings.

The different results derived from the different possible settings for a tank are shown below in Exhibit 49.

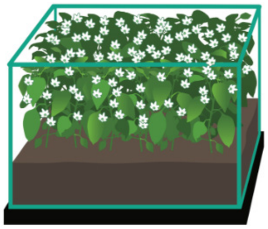
Exhibit 49: *Pepper Plants* Screen 9 – Possible Results for a Growth Tank

A 20 ml
B 0 ml
💧 300 or 600 ml



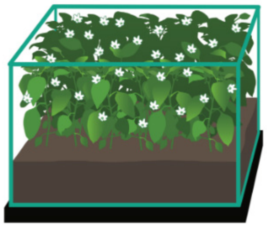
**Short Plants,
Many Flowers**

A 10 ml
B 0 ml
💧 300 or 600 ml



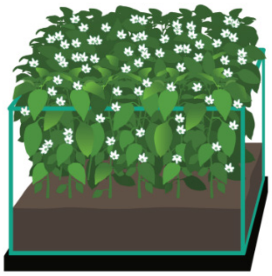
**Short Plants,
Medium Flowers**

A 0 ml
B 0 ml
💧 300 or 600 ml



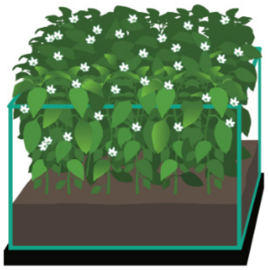
**Short Plants,
Few Flowers**

A 10 or 20 ml
B 10 or 20 ml
💧 300 or 600 ml



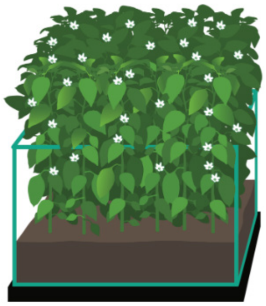
**Medium Plants,
Medium Flowers**

A 0 ml
B 10 ml
💧 300 or 600 ml



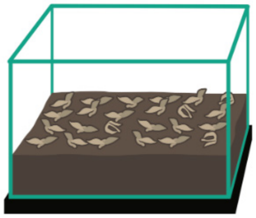
**Medium Plants,
Few Flowers**

A 0 ml
B 20 ml
💧 300 or 600 ml



**Tall Plants,
Few Flowers**

A 0, 10, or 20 ml
B 0, 10, or 20 ml
💧 0 ml



Dead Plants

Exhibit 50 shows the percentage of students responding correctly to Screen 9. A correct response was based on a correct experimental setup (from Exhibit 43, Screen 4) and the student selecting “Fertilizer A” as the better fertilizer because it produced the largest number of flowers. As an authentic activity, it was interesting to ask students to interpret the results of their own experiment. However, it was not a good measure of their interpretive skills because this depended on having designed a fully correct experiment—something achieved only by 15 percent of the students on average across countries (see Exhibit 43). As might have been anticipated, nearly all of the students who designed a fully correct experiment were able to interpret their own results (13% on average across countries). On average girls had somewhat higher performance than boys (14% vs. 12%).

Students who designed an incorrect experiment were unable to interpret their results in terms of the best fertilizer, but could select: “The results do not show which fertilizer is best for producing peppers.” Recognition of this also was “correct” in the sense that it was accurate, although these students had a different task than the students with correct results, and a number of different tasks among them depending on the particular flaws in their designs. On average across countries, 19 percent of students had incorrect experimental designs, and correctly selected not being able to say which was the best pepper producing fertilizer, with a correct explanation of why and a few more responded the same way based on incomplete experiments (2%). Larger percentages of students by far in the Russian Federation (65%) and the United States (58%) were able to recognize and describe the results of their incorrect setups as inconclusive than students in the rest of the countries (21% or fewer).

Exhibit 50

Pepper Plants Screen 9 – Percent Correct Overall and by Gender

Country	Percent Correct Fair Test and Control and Correct Interpretation (Fertilizer A with Explanation of Most Flowers)			Percent Correct Interpretation for Inconclusive Results	
	Overall Country	Girls	Boys	Incorrect Setup	Missing Values
Singapore	42 (2.0)	43 (3.0)	41 (2.6)	13 (1.2)	7 (1.2)
Chinese Taipei	28 (2.2)	29 (2.9)	27 (2.9)	18 (1.7)	2 (0.5)
Malaysia	19 (1.3)	21 (1.9)	18 (1.7)	16 (1.0)	1 (0.4)
Hong Kong SAR	17 (2.0)	20 (3.2)	14 (2.8)	11 (1.8)	0 (0.0)
Georgia	15 (2.2)	16 (2.9)	14 (3.1)	3 (1.0)	7 (1.3)
United States	14 (1.4)	16 (1.9)	13 (1.7)	58 (1.8)	2 (0.4)
Turkey	14 (1.6)	13 (1.8)	15 (2.4)	2 (0.6)	3 (0.8)
Finland	13 (1.4)	15 (2.1)	12 (2.0)	21 (1.5)	1 (0.3)
Sweden	13 (1.9)	14 (2.6)	12 (2.1)	18 (1.6)	2 (0.6)
England	13 (1.9)	14 (2.5)	12 (2.5)	20 (1.9)	2 (0.8)
Korea, Rep. of	11 (1.2)	15 (2.2)	8 (1.7)	21 (2.2)	1 (0.5)
Russian Federation	11 (1.4)	13 (2.3)	9 (1.6)	65 (2.2)	4 (1.0)
Israel	11 (1.5)	11 (2.4)	10 (2.3)	21 (1.9)	1 (0.5)
Norway (9)	10 (1.5)	12 (2.3)	8 (1.7)	17 (1.8)	4 (1.0)
Lithuania	10 (1.4)	13 (2.2)	7 (1.8)	12 (1.6)	1 (0.5)
Portugal	7 (1.6)	5 (1.4)	10 (2.5)	17 (1.9)	0 (0.3)
United Arab Emirates	7 (0.6)	8 (0.7)	7 (1.0)	9 (0.5)	1 (0.2)
Hungary	6 (0.9)	5 (1.3)	7 (1.5)	20 (1.4)	1 (0.5)
Italy	5 (1.0)	6 (1.5)	4 (1.4)	15 (1.7)	0 (0.2)
Qatar	4 (1.3)	5 (1.4)	3 (1.8)	6 (1.3)	0 (0.1)
France	3 (0.8)	4 (1.3)	3 (1.0)	20 (1.6)	3 (0.9)
Chile	2 (0.5)	2 (0.8)	2 (0.9)	16 (1.8)	0 (0.1)
International Average	**13 (0.3)**	**14 (0.5)** ▲	**12 (0.4)**	**19 (0.3)**	**2 (0.1)**
Benchmarking Participants					
Moscow City, Russian Fed.	21 (2.2)	18 (2.8)	25 (3.0)	14 (1.8)	3 (0.7)
Dubai, UAE	17 (1.6)	19 (2.1)	16 (2.2)	11 (1.2)	3 (0.6)
Ontario, Canada	13 (2.0)	16 (2.5)	10 (2.6)	21 (2.0)	1 (0.5)
Quebec, Canada	11 (1.4)	14 (2.3)	7 (1.7)	24 (2.0)	6 (1.0)
Abu Dhabi, UAE	3 (0.7)	3 (0.9)	3 (0.7)	8 (0.9)	1 (0.2)

▲ Percentage significantly higher than other gender

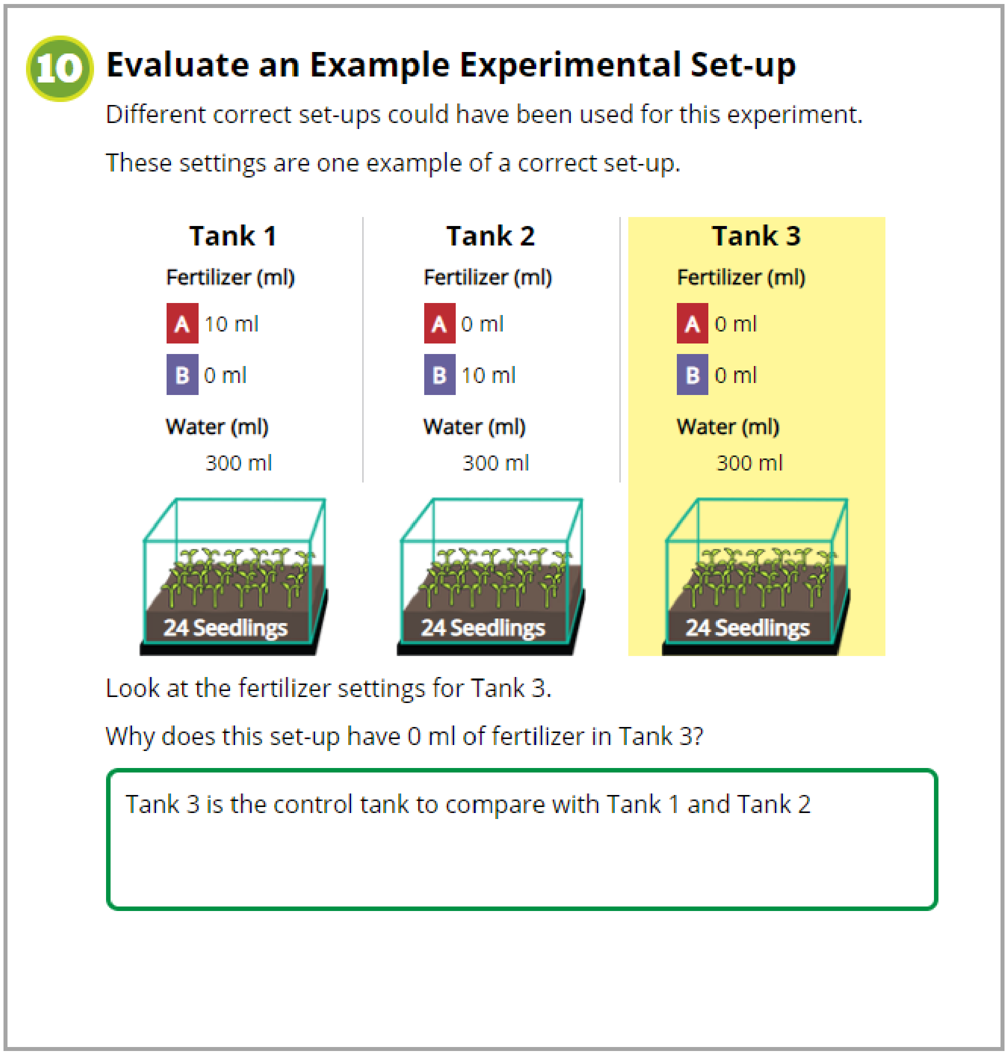
( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 10 – Evaluate an Example Experimental Setup

Screens 10 through 13 assessed students' understanding of a correct experimental setup and their performance in evaluating an experiment's results. Screens 10 through 13 are based on new, different examples of correct and incorrect experiments to make these items fair to all students (i.e., independent of the results of their experimental set up).

Results from the TIMSS 2019 field test indicated that use of a control group was one area of experimental design that was difficult for many students, and so Screen 10 was developed to probe directly into this issue. This screen presents an example of a correct experimental setup with a control tank (Tank 3) and asks students to explain why there is no fertilizer in Tank 3. As shown below, correct responses explained that a control tank provides a basis for comparing the effectiveness of each fertilizer.



10 Evaluate an Example Experimental Set-up

Different correct set-ups could have been used for this experiment.

These settings are one example of a correct set-up.

	Tank 1	Tank 2	Tank 3
Fertilizer (ml)			
A	10 ml	0 ml	0 ml
B	0 ml	10 ml	0 ml
Water (ml)	300 ml	300 ml	300 ml

Look at the fertilizer settings for Tank 3.

Why does this set-up have 0 ml of fertilizer in Tank 3?

Tank 3 is the control tank to compare with Tank 1 and Tank 2

Maximum Score Points: 1
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Reasoning
Science Practice: Generating Evidence

Exhibit 51 shows that although 65 percent of Singaporean students answered correctly, 41 percent or fewer did so in the remaining countries. On average across countries, just 23 percent of students answered correctly. Across countries, girls performed better than boys by a slight margin on average (24% vs. 22%).

Exhibit 51

Pepper Plants Screen 10 – Percent Correct Overall and by Gender

Country	Percent Correct (Explains that Tank 3 is the Control Tank)		
	Overall Country	Girls	Boys
Singapore	65 (2.1)	65 (2.7)	65 (2.8)
United States	41 (2.1)	42 (2.9)	40 (2.6)
Israel	38 (2.2)	39 (3.0)	37 (3.2)
Finland	37 (1.7)	45 (2.8)	31 (2.5)
Korea, Rep. of	35 (2.1)	40 (2.9)	31 (3.1)
Chinese Taipei	32 (2.0)	31 (3.3)	32 (2.6)
Lithuania	31 (2.2)	32 (3.3)	30 (3.0)
Hong Kong SAR	29 (3.0)	28 (3.6)	31 (4.1)
Sweden	27 (2.4)	30 (3.3)	24 (3.1)
England	21 (2.1)	19 (2.6)	23 (3.0)
United Arab Emirates	18 (0.8)	18 (1.2)	18 (1.2)
Turkey	17 (1.7)	16 (2.1)	18 (2.3)
Italy	16 (1.7)	21 (2.8)	12 (1.8)
Russian Federation	15 (1.5)	14 (2.2)	16 (2.1)
Malaysia	13 (1.5)	14 (1.8)	12 (1.9)
France	12 (1.5)	14 (2.1)	10 (1.8)
Norway (9)	12 (1.9)	13 (2.8)	11 (2.8)
Portugal	11 (1.6)	10 (1.8)	11 (2.4)
Hungary	10 (1.4)	10 (2.0)	10 (1.8)
Chile	10 (1.3)	11 (2.0)	9 (1.9)
Qatar	10 (1.4)	11 (2.0)	8 (2.4)
Georgia	8 (1.7)	10 (2.7)	6 (1.9)
International Average	**23 (0.4)**	**24 (0.6)** ▲	**22 (0.5)**
Benchmarking Participants			
Quebec, Canada	37 (3.2)	38 (4.1)	35 (3.9)
Moscow City, Russian Fed.	36 (2.1)	35 (3.0)	37 (2.9)
Dubai, UAE	33 (2.1)	31 (2.8)	35 (2.6)
Ontario, Canada	26 (2.4)	33 (3.8)	20 (2.9)
Abu Dhabi, UAE	13 (1.1)	14 (2.2)	11 (1.5)

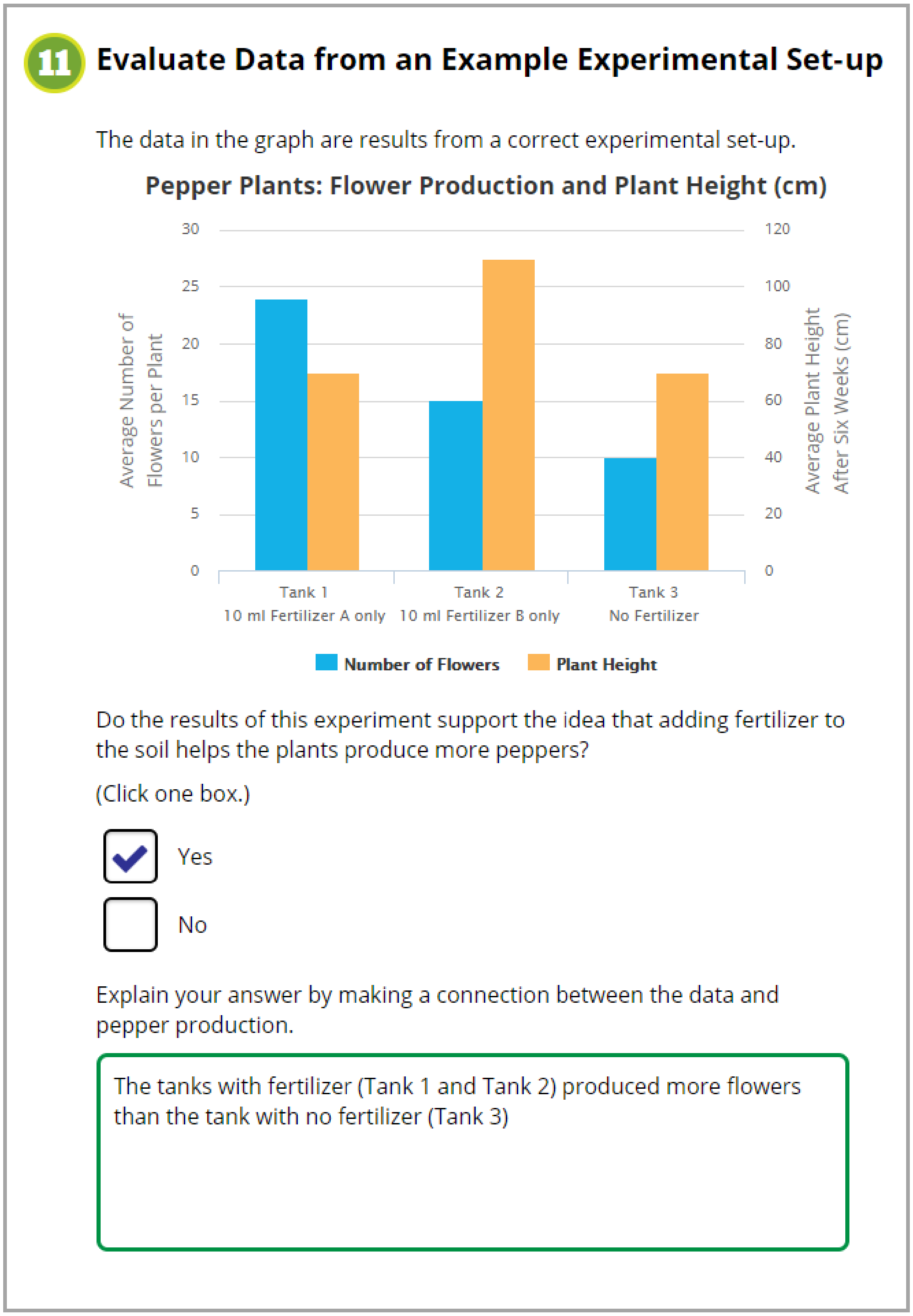
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 11 – Evaluate Data from an Example Experimental Setup

Screen 11 displays the results of a correct experimental setup, providing all students with the opportunity to interpret the same data. Correct responses indicated that the graph showed fertilizer helped pepper plant production because both tanks with fertilizer produced more flowers (and consequently would produce more peppers) than the tank with no fertilizer.



11 Evaluate Data from an Example Experimental Set-up

The data in the graph are results from a correct experimental set-up.

Pepper Plants: Flower Production and Plant Height (cm)

Do the results of this experiment support the idea that adding fertilizer to the soil helps the plants produce more peppers?

(Click one box.)

☑ Yes

□ No

Explain your answer by making a connection between the data and pepper production.

The tanks with fertilizer (Tank 1 and Tank 2) produced more flowers than the tank with no fertilizer (Tank 3)

Maximum Score Points: 1
Content Domain: Biology
Topic Area: Life Cycles, Reproduction, and Heredity
Cognitive Domain: Applying
Science Practice: Working with Data

Exhibit 52 shows the average percent correct across countries was 21 percent. Some of the explanations judged incorrect did include information from the graph. For example, 5 percent on average mentioned the two tanks with flowers but not the third tank, and a few, 2 percent on average, mentioned the height of the plants. In general, however, the eighth grade students had difficulty evaluating the results shown in the graph. Girls outperformed boys on average across countries.

Exhibit 52

Pepper Plants Screen 11 – Percent Correct Overall and by Gender

Country	Percent Correct (Gives Explanation Comparing the Number of Flowers in All 3 Tanks)			Percent Comparing the Number of Flowers in Only 2 Tanks (Incorrect)	Percent Comparing the Height of the Plants (Incorrect)
	Overall Country	Girls	Boys		
Singapore	50 (2.0)	50 (2.7)	50 (2.9)	7 (1.1)	2 (0.5)
Finland	32 (1.8)	39 (2.7)	24 (2.3)	1 (0.4)	2 (0.5)
Israel	30 (2.5)	32 (3.8)	28 (3.3)	9 (1.2)	4 (0.8)
Chinese Taipei	30 (1.8)	32 (2.8)	28 (2.5)	8 (1.0)	3 (0.8)
Qatar	28 (1.9)	28 (2.8)	27 (2.5)	0 (0.4)	2 (0.5)
United States	28 (1.8)	29 (2.2)	26 (2.2)	6 (0.6)	3 (0.5)
France	27 (1.9)	28 (3.2)	25 (2.7)	8 (1.2)	7 (1.1)
Korea, Rep. of	27 (1.9)	34 (3.4)	21 (2.4)	3 (0.7)	6 (1.1)
Turkey	24 (2.2)	26 (3.7)	20 (2.4)	2 (0.6)	2 (0.6)
Hong Kong SAR	24 (2.9)	24 (4.2)	23 (3.4)	7 (1.2)	2 (0.5)
Sweden	23 (2.1)	24 (3.2)	23 (3.0)	8 (1.3)	0 (0.1)
Hungary	22 (1.9)	25 (2.7)	19 (2.2)	7 (1.3)	2 (0.7)
England	20 (2.1)	22 (2.9)	18 (2.8)	4 (0.9)	2 (0.6)
Portugal	20 (1.8)	19 (3.0)	20 (2.5)	11 (1.6)	3 (0.9)
Norway (9)	18 (1.9)	18 (2.5)	17 (2.5)	1 (0.4)	3 (0.9)
Lithuania	15 (1.8)	20 (2.7)	10 (2.3)	6 (1.1)	1 (0.5)
Italy	15 (1.7)	16 (2.5)	14 (2.3)	2 (0.8)	0 (0.3)
Russian Federation	11 (1.5)	9 (1.6)	13 (2.1)	8 (1.5)	1 (0.3)
United Arab Emirates	10 (0.6)	12 (1.1)	8 (0.7)	3 (0.4)	2 (0.3)
Chile	6 (1.2)	5 (1.4)	6 (1.8)	0 (0.3)	4 (0.7)
Malaysia	5 (0.8)	7 (1.4)	3 (0.8)	1 (0.2)	1 (0.3)
Georgia	3 (1.1)	5 (2.0)	1 (0.8)	11 (1.8)	2 (0.5)
International Average	**21 (0.4)**	**23 (0.6)** ▲	**19 (0.5)**	**5 (0.2)**	**2 (0.1)**
Benchmarking Participants					
Moscow City, Russian Fed.	28 (2.3)	29 (3.4)	28 (3.0)	9 (1.3)	0 (0.2)
Ontario, Canada	27 (2.4)	29 (4.0)	24 (2.9)	5 (1.0)	3 (0.8)
Quebec, Canada	21 (2.1)	26 (3.0)	17 (2.7)	6 (1.0)	7 (1.3)
Dubai, UAE	19 (1.6)	21 (3.1)	17 (1.6)	5 (0.9)	2 (0.6)
Abu Dhabi, UAE	5 (0.9)	6 (1.3)	4 (1.0)	3 (0.6)	3 (0.5)

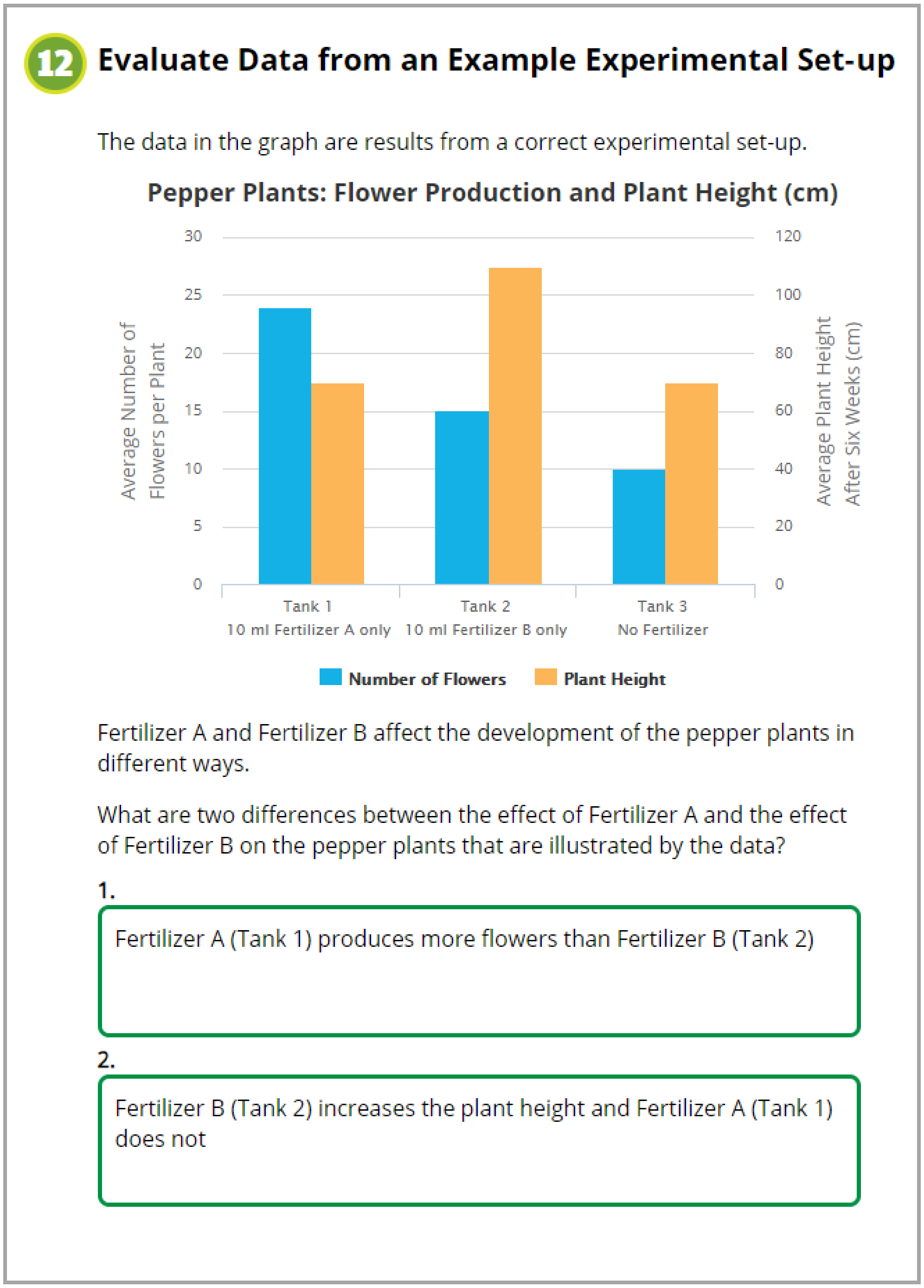
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 12 – Evaluate Data from an Example Experimental Setup

Screen 12 presents the same graph as shown in Screen 11, but this item asks students to describe the effects of Fertilizer A on pepper plant growth compared to the effects of Fertilizer B. Students were awarded full credit (2 points) if their responses included two elements: 1) compared to no fertilizer, Fertilizer A increased the number of flowers in the pepper plants more than Fertilizer B did and 2) Fertilizer B affected the height of the plants while Fertilizer A did not. Students whose responses only included one of these elements received partial credit (1 point).



12 Evaluate Data from an Example Experimental Set-up

The data in the graph are results from a correct experimental set-up.

Pepper Plants: Flower Production and Plant Height (cm)

Fertilizer A and Fertilizer B affect the development of the pepper plants in different ways.

What are two differences between the effect of Fertilizer A and the effect of Fertilizer B on the pepper plants that are illustrated by the data?

1.

Fertilizer A (Tank 1) produces more flowers than Fertilizer B (Tank 2)

2.

Fertilizer B (Tank 2) increases the plant height and Fertilizer A (Tank 1) does not

Maximum Score Points: 2
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Applying
Science Practice: Working with Data

Exhibit 53 shows 38 percent of students on average across countries having earned full credit by giving two differences between the effects of the fertilizers. Another 16 percent on average earned partial credit by giving one difference. Girls performed better than boys, on average across countries.

Exhibit 53

Pepper Plants Screen 12 – Percent Full Credit Overall and by Gender

Country	Percent Full Credit (Gives 2 Correct Differences Between Fertilizers)			Percent Partial Credit (Gives Only 1 Correct Difference)
	Overall Country	Girls	Boys	
Singapore	62 (1.6)	62 (2.4)	62 (2.5)	19 (1.4)
United States	54 (2.2)	56 (3.0)	53 (2.5)	16 (1.7)
Sweden	54 (2.5)	58 (3.5)	50 (3.8)	6 (1.0)
Hungary	53 (1.9)	55 (3.2)	51 (2.7)	9 (1.5)
France	53 (2.3)	56 (3.2)	50 (3.7)	12 (1.3)
Finland	50 (2.0)	55 (2.7)	46 (2.9)	15 (1.3)
England	48 (2.6)	50 (3.5)	46 (3.6)	10 (1.6)
Malaysia	48 (2.2)	52 (2.8)	44 (2.7)	12 (1.3)
Turkey	46 (2.3)	46 (3.3)	46 (3.8)	7 (1.0)
Portugal	39 (2.3)	36 (3.0)	42 (3.7)	19 (2.0)
Korea, Rep. of	39 (2.3)	45 (3.4)	33 (3.0)	26 (1.9)
Georgia	37 (2.9)	39 (4.0)	35 (3.5)	6 (1.1)
United Arab Emirates	37 (1.0)	41 (1.4)	34 (1.5)	9 (0.7)
Qatar	34 (2.4)	37 (3.8)	32 (2.5)	8 (1.4)
Chinese Taipei	34 (1.7)	36 (2.6)	32 (2.3)	6 (1.0)
Chile	34 (2.4)	35 (2.6)	32 (4.3)	22 (2.4)
Hong Kong SAR	25 (2.7)	29 (4.4)	22 (2.9)	7 (1.4)
Israel	22 (2.1)	22 (2.7)	22 (3.1)	21 (1.9)
Lithuania	21 (1.8)	26 (2.9)	17 (2.7)	40 (2.4)
Russian Federation	20 (2.3)	22 (3.3)	17 (2.3)	17 (1.8)
Norway (9)	14 (1.7)	16 (2.7)	13 (2.3)	37 (2.6)
Italy	11 (1.6)	14 (2.2)	9 (2.0)	32 (2.5)
International Average	**38 (0.5)**	**40 (0.7)** ▲	**36 (0.6)**	**16 (0.4)**
Benchmarking Participants				
Quebec, Canada	61 (2.5)	65 (3.8)	58 (3.9)	13 (1.6)
Dubai, UAE	58 (1.7)	60 (2.4)	55 (3.2)	12 (1.6)
Ontario, Canada	56 (2.2)	60 (3.7)	53 (3.3)	16 (1.9)
Moscow City, Russian Fed.	51 (2.3)	50 (3.2)	51 (3.1)	12 (1.7)
Abu Dhabi, UAE	29 (1.3)	33 (2.1)	25 (1.8)	7 (0.7)

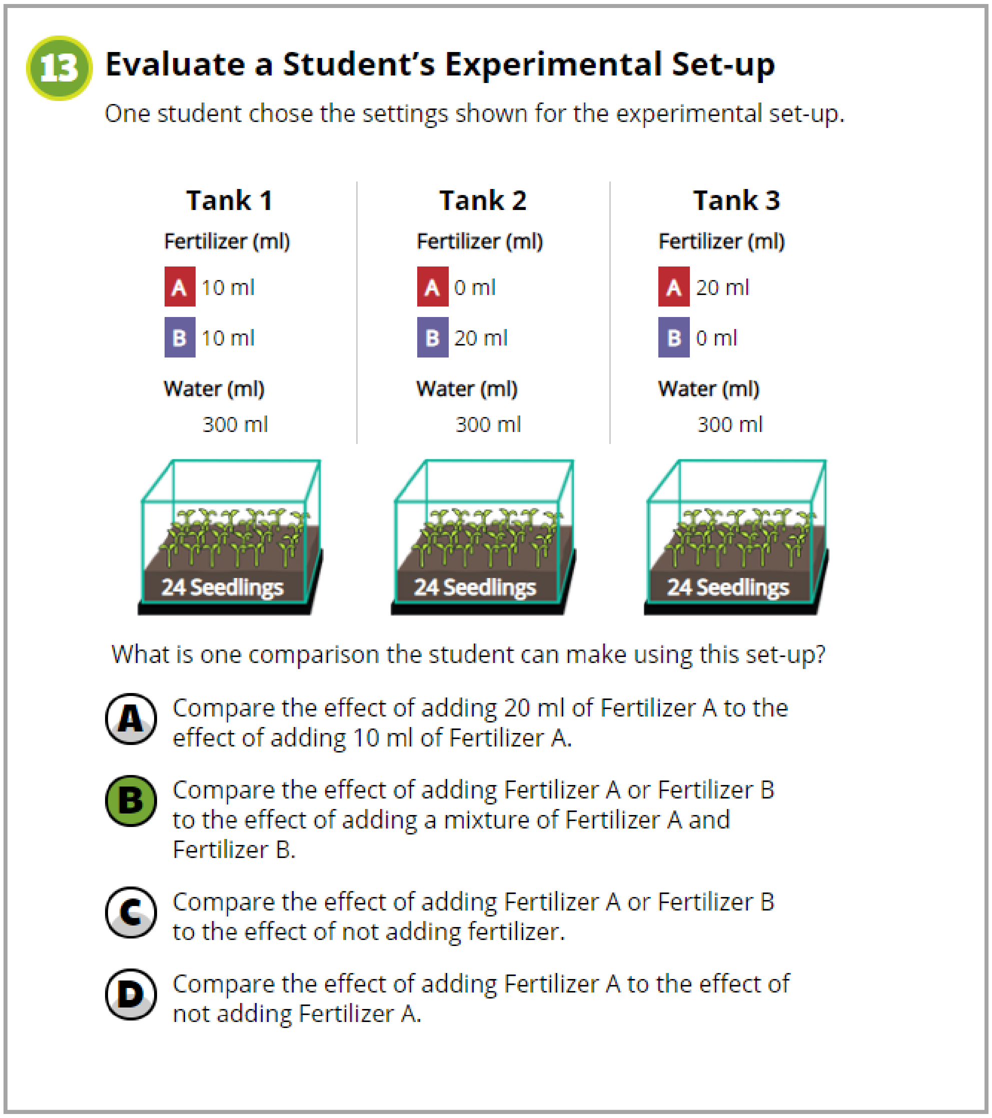
▲ Percentage significantly higher than other gender

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 13 – Evaluate a Student’s Experimental Setup

The final item in the *Pepper Plants* task addressed another aspect of students’ understanding of experimental designs. Students were given an experimental setup where three tanks have the same amount of fertilizer, but Tank 1 has half Fertilizer A and half B, Tank 2 only has Fertilizer B, and Tank 3 only has Fertilizer A. When asked what comparison could be made using the setup, the students needed to recognize the correct answer: the effect of adding Fertilizer A or Fertilizer B compared to the effect of adding a mixture of the two fertilizers (option B).



13 **Evaluate a Student’s Experimental Set-up**

One student chose the settings shown for the experimental set-up.

Tank 1	Tank 2	Tank 3
Fertilizer (ml)	Fertilizer (ml)	Fertilizer (ml)
A 10 ml	A 0 ml	A 20 ml
B 10 ml	B 20 ml	B 0 ml
Water (ml)	Water (ml)	Water (ml)
300 ml	300 ml	300 ml

What is one comparison the student can make using this set-up?

A Compare the effect of adding 20 ml of Fertilizer A to the effect of adding 10 ml of Fertilizer A.

B Compare the effect of adding Fertilizer A or Fertilizer B to the effect of adding a mixture of Fertilizer A and Fertilizer B.

C Compare the effect of adding Fertilizer A or Fertilizer B to the effect of not adding fertilizer.

D Compare the effect of adding Fertilizer A to the effect of not adding Fertilizer A.

Maximum Score Points: 1
Content Domain: Biology
Topic Area: Ecosystems
Cognitive Domain: Reasoning
Science Practice: Generating Evidence

Exhibit 54 shows that eighth grade students performed relatively well on this item, with 59 percent selecting the correct response on average across countries. There were essentially no performance differences between girls and boys, on average across countries.

Pepper Plants Screen 13 – Percent Correct Overall and by Gender

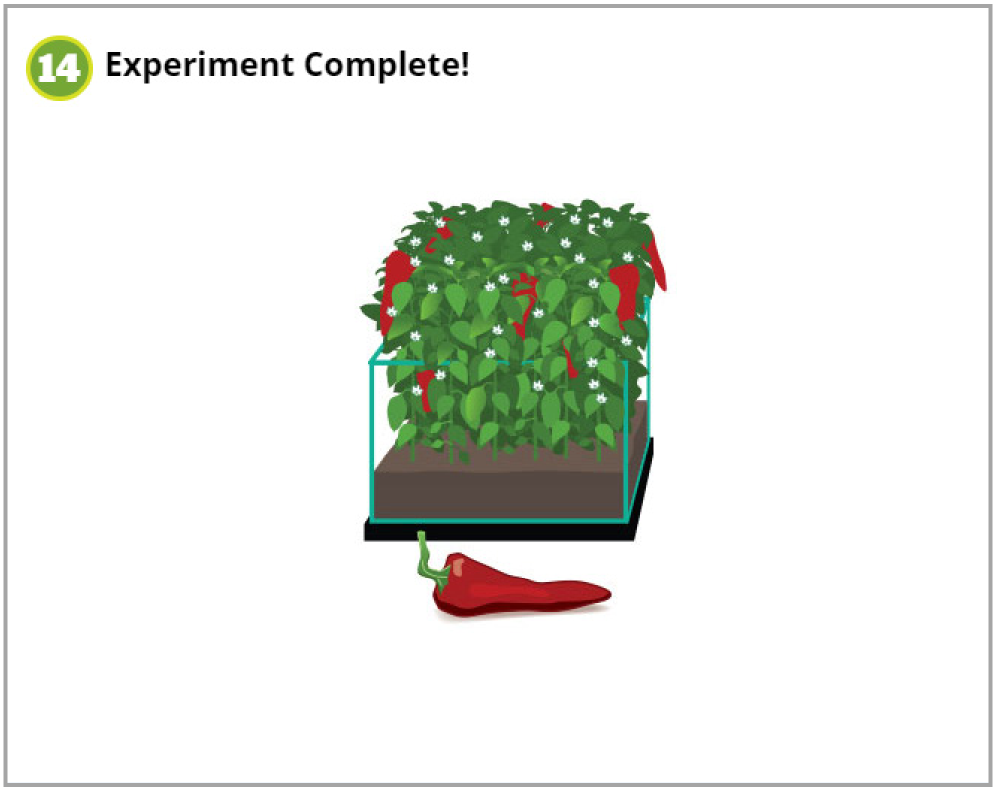
Country	Percent Correct (Compare the effect of adding Fertilizer A or Fertilizer B to the effect of adding a mixture of A and B)		
	Overall Country	Girls	Boys
Singapore	74 (1.5)	75 (2.1)	72 (2.3)
Chinese Taipei	70 (1.7)	69 (2.5)	70 (2.4)
Russian Federation	66 (2.1)	67 (3.1)	65 (2.6)
Portugal	66 (2.2)	61 (3.3)	70 (3.3)
Korea, Rep. of	65 (2.2)	67 (2.9)	62 (3.1)
Hong Kong SAR	64 (2.1)	62 (3.5)	66 (2.7)
Sweden	64 (2.3)	67 (3.4)	61 (4.1)
Israel	64 (2.5)	59 (3.2)	68 (3.5)
Finland	63 (1.9)	69 (2.3)	57 (3.0)
United States	62 (1.4)	60 (2.6)	64 (2.1)
Chile	61 (3.3)	64 (4.5)	58 (3.7)
Lithuania	60 (2.7)	63 (3.5)	57 (3.7)
Norway (9)	58 (2.3)	62 (3.5)	54 (3.5)
France	57 (2.5)	57 (3.4)	57 (3.5)
Hungary	56 (2.5)	53 (3.8)	60 (3.2)
Italy	55 (2.4)	56 (3.4)	55 (2.9)
England	52 (2.6)	55 (3.2)	49 (3.9)
Malaysia	52 (1.5)	49 (2.2)	54 (2.6)
United Arab Emirates	49 (1.2)	51 (1.6)	47 (1.4)
Turkey	49 (2.2)	44 (2.9)	53 (3.6)
Qatar	45 (3.0)	49 (4.7)	42 (3.4)
Georgia	38 (2.9)	36 (3.3)	40 (4.6)
International Average	**59 (0.5)**	**59 (0.7)**	**58 (0.7)**
Benchmarking Participants			
Moscow City, Russian Fed.	75 (1.9)	73 (2.8)	77 (2.7)
Ontario, Canada	64 (2.3)	71 (3.3)	59 (2.9)
Quebec, Canada	61 (2.9)	60 (3.8)	62 (3.6)
Dubai, UAE	57 (2.1)	60 (3.0)	53 (2.4)
Abu Dhabi, UAE	49 (1.8)	52 (2.4)	45 (2.2)

( ) Standard errors are shown in parentheses. Because of rounding some results may appear inconsistent.

SOURCE: IEA's Trends in International Mathematics and Science Study - TIMSS 2019

Screen 14 – Experiment Complete!

The last screen of the PSI task confirmed that the experiment was complete and no further items would be presented.



14 Experiment Complete!

Conclusion and Reflections

Pepper Plants began the TIMSS 2019 PSI task development process, and had the longest evolution and the most revisions. As an exercise in experimental design and interpretation in a biological context, great effort was expended in developing a scenario that was as realistic as possible (albeit a simulated situation) and that provided students with interesting and engaging activities in designing and setting up their experiment.

- *Pepper Plants* seems to have hit the right note in terms of student engagement and item difficulty, with very few students not managing to complete the task in the time allowed.
- Experience from the TIMSS field test confirmed that the activities involved in designing and setting up the experiment, such as deciding which fertilizer to apply and in what amount, and how much water to supply, were engaging and motivating for the students.

- However, the field test also showed that a well-balanced task would, in addition, require emphasis on students' reasons for their design choices, their understanding of the characteristics of good design, and their ability to interpret and make generalizations from the results. The final version included items addressing these issues.
- The *Pepper Plants* task shows that experimental setup and design can be a fruitful area for future Problem Solving and Inquiry tasks. The results indicate that eighth grade students have a reasonable grasp of the idea of an experiment as a fair test, but less understanding of an effective experimental control.

APPENDIX A

Problem Solving and Inquiry (PSI) Tasks in the TIMSS 2019 Assessment Design: A Look at Booklet Completion Rates

Basic Design of Blocks and Booklets in TIMSS 2019

The *TIMSS 2019 Assessment Design* encompassed eTIMSS and paperTIMSS as mirrored efforts of each other, such that within each, both fourth and eighth grades consisted of mathematics and science achievement items grouped into 28 item blocks (14 mathematics and 14 science).[1] The blocks were arranged into 14 booklets with four blocks per booklet. There were two parallel variations of the design: one for the paper version (paperTIMSS) and another adapted to digital presentation mode (eTIMSS). Because the special Problem Solving and Inquiry (PSI) tasks described herein are inherently computer based and had no paper counterpart, they were treated as a separate addition to the eTIMSS version.

Each TIMSS 2019 booklet or eBooklet consisted of two blocks of mathematics and two blocks of science items. Each item block appeared in two booklets, in a different position in each booklet to account for effects on achievement that can occur from items being earlier or later in the testing sessions (e.g., learning or fatigue). Each student completed one booklet. The 14 booklets were distributed among the students in participating classes according to a random assignment procedure, so that each booklet or eBooklet was assigned to approximately equal percentages of students.

Keeping Booklets 1–14 the same as possible in both paperTIMSS and eTIMSS enabled establishing a link between the paper and digital assessment modes, and made it possible to report student achievement on both paperTIMSS and eTIMSS together on the same TIMSS achievement scale in *TIMSS 2019 International Results in Mathematics and Science*.[2,3] These results which IEA released in December 2020 presented a comprehensive view internationally of mathematics and science achievement at the fourth and eighth grades as well as contextual data for 64 countries and 8 benchmarking systems.

The Problem Solving and Inquiry eBooklets

To explore how to increase the benefits of digital assessment for TIMSS 2023 (the subject of this report), at each grade, eTIMSS also included four blocks of problem solving and inquiry tasks and items, two of mathematics and two of science. Because the items in eBooklets 1–14 were the same as for paperTIMSS, the PSI item blocks were assigned to eBooklets 15–16. Similar to the paperTIMSS assessment, all 16 eBooklets were randomly distributed among students in the eTIMSS classes.

Allocating the PSI tasks to separate booklets (eBooklets 15–16) that were rotated along with eBooklets 1–14 provided the means to conduct a subsequent achievement scaling that also placed the PSI tasks on the TIMSS achievement scale.[4] As described in the Introduction to this report, there essentially was no difference in average eTIMSS achievement as a result of including the PSI data. Although the PSI initiative was important for looking forward to TIMSS 2023, the PSIs only represented a small percentage of the nearly 900 items in the TIMSS 2019 assessment and the students who took them comprised just 12 percent of the 600,000 students.

The PSI tasks and items were allocated to four assessment blocks as shown in Exhibit A.1.

Exhibit A.1: PSI Assessment Blocks with Subject and Task Label

4th Grade Blocks	Subject	PSI Task(s)
M1	Mathematics	Penguins + Robots-4 (Secure)
M2	Mathematics	School Party
S1	Science	Farm Investigation
S2	Science	Sugar Experiment (Secure)

8th Grade Blocks	Subject	PSI Task(s)
M1	Mathematics	Building + Robots-8
M2	Mathematics	Dinosaur Speed (Secure)
S1	Science	Sunken Ship (Secure)
S2	Science	Pepper Plants

Exhibit A.1 presents, for both fourth and eighth grades, the PSI item block labels from the TIMSS 2019 assessment design, the subjects that they address, and the PSI tasks that were assigned to them. For example, at fourth grade, block M1 is a mathematics block containing one longer task, *Penguins*, combined with a shorter *Robots-4* task, whereas S1 is a science block containing the single *Farm Investigation* PSI task. Four of the PSI blocks, *School Party* and *Farm Investigation* at fourth grade and *Building* + *Robots-8* and *Pepper Plants* at eighth grade, are published in full in this report. The other four blocks have been kept secure for use in future TIMSS assessments.

Exhibit A.2 shows, for both fourth and eighth grades, how the PSI assessment blocks were allocated to eBooklets 15 and 16 according to the eTIMSS 2019 assessment design. Each booklet has two mathematics and two science blocks, presented in two separately timed sessions. Blocks are not separately timed within sessions. At fourth grade, each session lasts for 36 minutes, with a 15-minute break in between. Eighth grade sessions last for 45 minutes.

Exhibit A.2: PSI Block Assignments to eBooklets*

4th Grade	36 minutes Session 1 Block Position 1 & 2	36 minutes Session 2 Block Position 3 & 4
eBooklet 15	Penguins + Robots-4 (M1); School Party (M2)	Farm Investigation (S1); Sugar Experiment (S2)
eBooklet 16	Sugar Experiment (S2); Farm Investigation (S1)	School Party (M2); Penguins + Robots-4 (M1)

8th Grade	45 minutes Session 1 Block Position 1 & 2	45 minutes Session 2 Block Position 3 & 4
eBooklet 15	Building + Robots-8 (M1); Dinosaur Speed (M2)	Sunken Ship (S1); Pepper Plants (S2)
eBooklet 16	Pepper Plants (S2); Sunken Ship (S1)	Dinosaur Speed (M2); Building + Robots-8 (M1)

* Students had a 15-minute break between Session 1 and Session 2.

Note that eBooklet 15 and eBooklet 16 contain the same assessment material, but presented in different orders. For example, at fourth grade, eBooklet 15 contains the two mathematics blocks, *Penguins + Robots-4* and *School Party* in the first session (Positions 1 & 2) and, then after the break, the two science blocks, *Farm Investigation* and *Sugar Experiment*, in the second session (Positions 3 & 4). eBooklet 16 begins with the two science blocks in the first session (Positions 1 & 2) but in the reverse order from eBooklet 15: *Sugar Experiment* followed by *Farm Investigation*. Session 2 (Positions 3 & 4) then contains the two mathematics blocks, but again in reverse order from eBooklet 15: *School Party* followed by *Penguins + Robots-4*. This block-booklet arrangement counterbalances the subject position effect (mathematics first, then science in eBooklet 15; science first, then mathematics in eBooklet 16). The arrangement also counterbalances the position-within-session effect to the extent possible (e.g., *Penguins + Robots-4* followed by *School Party* in Position 1 & 2 of the first session in eBooklet 15; *School Party* followed by *Penguins + Robots-4* in Position 3 & 4 of the second session).

Not Reached Items in the PSI Tasks

In reviewing the student response data for the PSIs, it became evident that students assigned PSI tasks (eBooklets 15 and 16) were not always completing their booklets at the same high rates as students assigned booklets containing regular, non-PSI items (eBooklets 1 to 14). As shown in Exhibit A.3, the percentage of students reaching all items in the regular eBooklets is very high, ranging from 92 percent for mathematics and science at fourth grade to 97 percent for eighth grade science. In contrast, students assigned PSI eBooklets had considerably lower completion rates, especially at fourth grade: 66 percent for mathematics and 76 percent for science.

Exhibit A.3: Student Completion Rates for Regular eTIMSS and PSI eBooklets

Percentage of Students Reaching All Items	4th Grade		8th Grade	
	Mathematics	Science	Mathematics	Science
Regular eTIMSS	92%	93%	94%	97%
PSIs	66%	76%	83%	94%

Because eBooklets 15 and 16 contain the same PSI blocks but in different orders, it was possible to investigate whether the lower completion rates were related to the block position in the booklet, or more specifically if a block presented earlier in a session had higher completion than a block later in a session. Exhibit A.4 addresses this issue by comparing the percentages of students not reaching the last item in each PSI block when the block is at the beginning and end of a session.

In eBooklet 15, the *Penguins + Robots-4* fourth grade mathematics task is in the first block position and students had no difficulty reaching the end of the task, with only 4 percent not reaching the last item. However, in eBooklet 16 this task is in Position 4, the last position at the end of the testing session, and the percentage of students not reaching the last item increased to 30 percent. A similar situation is apparent for *School Party*, the other fourth grade mathematics task. In eBooklet 16 this task is in Position 3, which is the first position in Session 2 after the break. Position 3 is similar to Position 1 in that students have just had a break and are starting afresh in a new, separately timed session. Students had no difficulty completing *School Party* in this position, with just 1 percent not reaching the last item in the block. In contrast, *School Party* is in block Position 2 in eBooklet 15, which is the end of the first session. Students assigned this booklet would have worked through the *Penguins + Robots-4* task before beginning *School Party*, and by the end of the task the percentage not reaching the last item increased to 39 percent.

Exhibit A.4: Percentage of Students Not Reaching the Last Item in each PSI Block, by Block Position*

4th Grade	36 minutes Session 1		36 minutes Session 2	
	Block Position 1	Block Position 2	Block Position 3	Block Position 4
eBooklet 15	Penguins + Robots-4 4%	School Party 39%	Farm Investigation 0%	Sugar Experiment 19%
eBooklet 16	Sugar Experiment 3%	Farm Investigation 28%	School Party 1%	Penguins + Robots-4 30%

8th Grade	45 minutes Session 1		45 minutes Session 2	
	Block Position 1	Block Position 2	Block Position 3	Block Position 4
eBooklet 15	Building + Robots-8 1%	Dinosaur Speed 14%	Sunken Ship 0%	Pepper Plants 5%
eBooklet 16	Pepper Plants 0%	Sunken Ship 8%	Dinosaur Speed 1%	Building + Robots-8 21%

* Students had a 15-minute break between Session 1 and Session 2.

The two fourth grade science PSI tasks show a block position effect similar to mathematics, although the percentage not reaching the last item is somewhat less: 19 and 28 percent for *Sugar Experiment* and *Farm Investigation*, respectively, compared with 30 and 39 percent for *Penguins + Robots-4* and *School Party*.

The completion rates for eighth grade were somewhat better than for fourth grade, as shown in Exhibit A.3, and this was reflected in lower percentages of students failing to reach the last item in a block (Exhibit A.4). The two mathematics PSI blocks (*Building* + *Robots-8* and *Dinosaur Speed*) had essentially no students not reaching the last item when these blocks were in the first position of a session (Position 1 and Position 3), but somewhat more (21% and 14%) when in the second session position (Position 2 and Position 4). The two eighth grade science PSI blocks showed the least not-reached effect, with essentially all students reaching all items when these blocks were in the first position of a session, and just 5 percent not reaching the last item in *Pepper Plants* and 8 percent in *Sunken Ship* when in the second position of a session.

Percent Correct in the PSI Report

Considering the relatively large percentages of students not reaching all of the items in the PSI blocks, it was clear that not all students had an opportunity to answer all of the items and that the usual TIMSS practice of treating not reached item responses as incorrect could introduce bias into the reported results. To avoid this situation, all not-reached responses were considered to be “not administered” rather than as incorrect, and not included in the computation of percent correct statistics. Accordingly, the base of the percent correct statistics presented in this report is the number of students that reached the item and had an opportunity to answer, rather than the number of students to whom the booklet was administered. The procedure of treating not-reached responses as not administered also was adopted in scaling the PSI data and scoring student responses.[5]

Notes

[1] Martin, M. O., Mullis, I. V. S., & Foy, P. (2017). TIMSS 2019 assessment design. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/frameworks/framework-chapters/assessment-design/

[2] Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

[3] Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 12.1–12.146). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html

[4] Fishbein, B., & Foy, P. (2021). Scaling the TIMSS 2019 problem solving and inquiry data. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 17.1–17.51). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-17.html

[5] Ibid.

APPENDIX B

Using Timing Data to Investigate Non-Response in the TIMSS 2019 Problem Solving and Inquiry (PSI) Tasks

Overview

Timing data collected as part of eTIMSS 2019 can be used to learn more about the patterns of non-response and block position effects described in Appendix A. Considering the relatively large percentages of students not reaching all of the items in the assessment eBooklets containing Problem Solving and Inquiry (PSI) tasks, it was clear that not all students had an opportunity to answer all of the items. However, reviewing the response data for the PSI tasks together with timing data indicated that instead of running out of time, some students stopped responding to items with plenty of time remaining, perhaps through fatigue or lack of motivation.

To explore this possibility, an investigation was conducted using measures derived from event log data collected as part of eTIMSS 2019. Event log data provide a comprehensive sequence of students' interactions with the computer-based assessment. A timestamp (in milliseconds) was saved for each student-object interaction, providing a full history of what the student clicked, entered, and selected. From the log data, information could be derived about students' test-taking behaviors, including response time, item visit and revisit behavior, and response revisions. Analyzing this information alongside the response data provided additional insights into student patterns of non-response.

Two phases of analysis were conducted: an item-level analysis and a student-level analysis. Analyses were conducted separately for each of fourth and eighth grades and for mathematics and science at the booklet level, focusing on the last item of each subject part (session). First, examining average timing measures for each PSI item showed that some students who did not reach all items actually had time remaining after giving their last response, but stopped responding to subsequent items. Then, to determine how many students stopped responding before time was up versus ran out of time, students who did not reach all items were classified into groups based on the time they gave their last response as a proxy for their last meaningful interaction with the items. The results for PSI students were compared to the results for students who took "regular" (non-PSI) eTIMSS items and indicated that the stopping behavior was much more common in the PSI tasks, particularly in mathematics, and was associated with lower performance on the PSIs. In mathematics, relatively more students stopped responding than ran out of time. In science, smaller percentages of students in each of the two groups were about equal.

Definition of Not Reached Items

The TIMSS definition of "Not Reached" assumes that students progress through assessment booklets, or eBooklets, in sequential order. Following TIMSS' standard data cleaning procedures developed for paper-based administration, if a student omitted two items in a row and all subsequent items in the booklet half were also blank, the second omitted item and all subsequent were coded in the data as "Not Reached." In the paper-based environment, there is no way to know if the student was working on the first omitted item, and so it was considered to be omitted rather than not reached. However, with the new information available through event log data, it was possible to determine whether a student actually visited an item screen, the time they arrived, and how long they spent. The analyses conducted for this Appendix take advantage of this information to make informed inferences about whether students did not reach all items because they ran out of time or because they stopped responding, perhaps because of fatigue or lack of motivation.

Item-Level Analysis

The item-level timing analysis revealed evidence that there were at least some students who stopped responding to PSI items sometime before the end of each session or subject part. This was true at both fourth and eighth grades and particularly in mathematics. Students were given the same amount of time to complete each of two sessions—36 minutes at the fourth grade and 45 minutes at the eighth grade. Students were given two blocks of mathematics PSI items in one session and two blocks of science PSI items in the other session, with a 15-minute break in between.

The item-level analysis indicated for each PSI item the percentage of students who were coded as not reaching the item, but who had a record of arriving on the screen containing the item. Timing averages for these students were compared to averages for students who did reach the item. This included 1) the average time that students arrived on the item screen (in minutes from 0—the start of the session), and 2) the average time that students spent on the screen (in minutes).

Exhibit B.1 presents the results for the last mathematics item in each PSI booklet at the fourth grade. As described in Appendix A, eBooklet 15 included two blocks of mathematics PSIs in the first half of the assessment—*Penguins* + *Robots-4*, followed by *School Party*. eBooklet 16 had *School Party* then *Penguins* + *Robots-4* in the second half. Fourth grade students had 36 minutes to complete the two PSI blocks.

On average across countries, students who reached the last School Party item in eBooklet 15 arrived at 27.3 minutes, leaving 8.7 minutes remaining until the end of the session (at 36 minutes).

Students who reached this item spent 1.84 minutes responding to the item, on average. Although 39 percent of students had the last item coded as "not reached," almost half of these students (16%) visited the screen at least once. The 16 percent of students with the item not reached arrived at 30.5 minutes with 5.5 minutes remaining, suggesting that these students arrived with plenty of time left, but did not respond. The small amount of time spent by the not-reached students (0.75 minutes) could be due to logging out of the test early or going back through previous screens.

Exhibit B.1: Timing Averages for the Last Mathematics PSI Item by eBooklet—Grade 4

eBooklet	Last Item Reached		Last Item Not Reached			
	Arrival Time (minutes)	Time Spent (minutes)	Total Percent	Percent Visited Screen	Arrival Time (minutes)	Time Spent (minutes)
eBooklet 15 (Positions 1 & 2)	27.3	1.84	39%	16%	30.5	0.75
eBooklet 16 (Positions 3 & 4)	25.1	2.48	30%	17%	25.2	1.50

Fourth grade students had 36 minutes to complete two blocks of mathematics PSIs.
Timing information is limited to the screen level. Percentages are at the item level.

The results for eBooklet 16 were similar, but with relatively fewer students failing to reach the last item (30% compared to 39% in eBooklet 15). On average, students who reached the last *Robots-4* item arrived at 25.1 minutes with 10.9 minutes remaining, and spent 2.48 minutes on the screen. More than half of the students who did not reach the item (17%) actually had been recorded as arriving on the screen at 25.2 minutes with 10.8 minutes remaining, on average. These students had much more time remaining than the average time spent by students who responded to the item (2.48 minutes).

Exhibit B.2 presents the results for fourth grade science. Fewer fourth grade students did not reach all the science items than did not reach all the mathematics items—19 percent in eBooklet 15 and 28 percent in eBooklet 16 (compared to 39% and 30% for mathematics, respectively). More students failed to reach all items (or did not respond to all items) when *Sugar Experiment* followed by *Farm Investigation* were in the first session than when Farm Investigation followed by *Sugar Experiment* were in the second session (28% in eBooklet 16 vs. 19% in eBooklet 15).

Exhibit B.2: Timing Averages for the Last Science PSI Item by eBooklet—Grade 4

eBooklet	Last Item Reached		Last Item Not Reached			
	Arrival Time (minutes)	Time Spent (minutes)	Total Percent	Percent Visited Screen	Arrival Time (minutes)	Time Spent (minutes)
eBooklet 15 (Positions 3 & 4)	25.6	0.59	19%	7%	24.8	0.13
eBooklet 16 (Positions 1 & 2)	27.3	1.19	28%	3%	30.8	0.13

Fourth grade students had 36 minutes to complete two blocks of science PSIs.
Timing information is limited to the screen level. Percentages are at the item level.

As reported in Exhibit B.2, students who reached the last science item in eBooklet 15 arrived at 25.6 minutes and in eBooklet 16 arrived at 27.3 minutes, on average. There were small percentages of students with the last item coded as “not reached” that actually visited the screen—7 percent in eBooklet 15 and 3 percent in eBooklet 16. However, given that students who reached the last item spent only 0.59–1.19 minutes on average, the students in question seem to have had ample time remaining to respond, arriving with 11.2 minutes remaining (eBooklet 15) and 5.2 minutes remaining (eBooklet 16), on average.

At the eighth grade, students had 45 minutes to complete each half of their booklets, but had more items to answer compared to fourth grade. Exhibit B.3 presents the results for the last eighth grade mathematics PSI items in eBooklets 15 and 16, respectively. In both booklets, the majority of students with not-reached codes for the last item had record of visiting the item screen—9 percent in eBooklet 15 (compared to 14% total not reached) and 18 percent in eBooklet 16 (compared to 21% total not reached), on average. These students arrived with 14.0–16.3 minutes remaining on the clock, on average. Students who reached the item spent 1.99–2.70 minutes responding on average, suggesting there was enough time remaining for at least some of the not-reached group to respond.

Exhibit B.3: Timing Averages for the Last Mathematics PSI Item by eBooklet—Grade 8

eBooklet	Last Item Reached		Last Item Not Reached			
	Arrival Time (minutes)	Time Spent (minutes)	Total Percent	Percent Visited Screen	Arrival Time (minutes)	Time Spent (minutes)
eBooklet 15 (Positions 1 & 2)	30.2	1.99	14%	9%	31.0	0.48
eBooklet 16 (Positions 3 & 4)	27.6	2.70	21%	18%	28.7	1.39

Eighth grade students had 45 minutes to complete two blocks of mathematics PSIs.
Timing information is limited to the screen level. Percentages are at the item level.

Similar to the results at the fourth grade, the eighth grade mathematics results showed evidence that the ordering of the PSIs within the booklet had an effect on average completion rates. In eBooklet 16 with *Dinosaur Speed* followed by *Building + Robots-8* in the second half, more students failed to reached all items than in eBooklet 15 when *Building + Robots-8* followed by *Dinosaur Speed* were in the first half (21% vs. 14%).

Based on the results for eighth grade science (Exhibit B.4), it could be said that the eighth grade science PSIs were the most successful in keeping students engaged. In eBooklet 15, only 5 percent of students did not reach the last item, and in eBooklet 16 only 8 percent did not reach the last item. Only 3 percent of not-reached students in each booklet had record of visiting the last item screen.

Exhibit B.4: Timing Averages for the Last Science PSI Item by eBooklet—Grade 8

eBooklet	Last Item Reached		Last Item Not Reached			
	Arrival Time (minutes)	Time Spent (minutes)	Total Percent	Percent Visited Screen	Arrival Time (minutes)	Time Spent (minutes)
eBooklet 15 (Positions 3 & 4)	27.9	0.81	5%	3%	23.4	0.10
eBooklet 16 (Positions 1 & 2)	30.4	1.69	8%	3%	33.3	0.28

Eighth grade students had 45 minutes to complete two blocks of science PSIs.
Timing information is limited to the screen level. Percentages are at the item level.

Student-Level Analysis

The item-level analysis made evident that at least some of the students who did not reach all items had enough time remaining, but stopped responding before the time expired. The next step involved determining the relative proportions of students who exhibited this stopping behavior versus running out of time. Toward this end, a time was derived for each student to indicate when their last response was given (or revised) in each booklet half (minutes from 0—the start of the session). This measure for “time of last response” served as an approximation of the time that students last meaningfully interacted with an item during the subject session.

Unfortunately, it was not possible to determine the precise time that students logged out of the test session. Therefore, it was necessary to implement a decision rule for when students had finished work on the assessment. This analysis assumes that among students who did not reach all items, those who remained active and gave a response within 30 seconds of the maximum allotted time (36 minutes at the fourth grade; 45 minutes at the eighth grade) ran out of time. On the other hand, those who did not interact with any item within 30 seconds of the time limit were assumed to have stopped responding. This 30 second cutoff was chosen based on an analysis of the distribution of time of last response across countries.[1,2]

At the fourth and eighth grades and for each subject, all students were classified according to the procedure described above, including all PSI students as well as students who took regular eTIMSS eBooklets 1–14. First, students who reached all items were classified as "Reached All Items." Among the students remaining, those who gave their last response within 30 seconds of the time limit (more than 35.5 minutes at the fourth grade; more than 44.5 minutes at the eighth grade) were classified as "Ran Out of Time." The students who did not interact with any item within 30 seconds of the time limit were classified as "Stopped Responding."

Exhibit B.5 presents the results for fourth grade mathematics. On average across countries, only 3 percent of students who took regular eTIMSS "Ran Out of Time" and 5 percent "Stopped Responding," with the majority (92%) reaching all items. The 5 percent who stopped responding gave their last response at 29.6 minutes, with 6.4 minutes remaining, on average.

Exhibit B.5: Student Response Type Classifications for Mathematics—Grade 4

Student Response Group	Percent of Students	Time of Last Response (minutes)	Percent of Items Correct	Percent of Items Not Reached
Regular eTIMSS Mathematics				
1. Reached All Items	92%	25.8	49%	0%
2. Ran Out of Time	3%	35.9	40%	16%
3. Stopped Responding	5%	29.6	36%	15%
Total		**26.2**	**49%**	**1%**
PSI Mathematics				
1. Reached All Items	66%	29.0	40%	0%
2. Ran Out of Time	14%	35.9	43%	15%
3. Stopped Responding	21%	29.7	36%	13%
Total		**30.1**	**40%**	**5%**

All statistics were computed at the student level by country, then averaged across countries.
Because of rounding some results may appear inconsistent.

The results for students who took PSI booklets tell a different story, with fewer fourth grade students having "Reached All Items" (66% of PSI students compared to 92% of regular students). Among PSI students, a higher percentage of students were classified as "Stopped Responding" compared to "Ran Out of Time"—21 percent versus 14 percent. On average across countries, PSI students who reached all mathematics items gave their last response at 29.0 minutes, with 7.0 minutes remaining in the session. Similar to the students who took the regular eTIMSS items, PSI students who stopped responding gave their last response at 29.7 minutes, not interacting with any item for more than 6 minutes, on average. As could be expected, these students had the lowest performance (36% of items correct) compared to students who reached all items and students who ran out of time (40–43% of items correct).

Results for fourth grade science are shown in Exhibit B.6. Similar to mathematics, the majority of students who took regular eTIMSS (93%) reached all items with an average time of 25.5 minutes. On the other hand, just 76 percent of PSI students reached all science items, and with an average time of 28.1 minutes. In contrast with the mathematics results which showed that relatively more students stopped responding than ran out of time (21% vs. 14%), science had about equal amounts of students in the two categories—13 percent and 11 percent, respectively. On average, students who stopped responding gave their last response at 30.6 minutes, with more than 5 minutes remaining. Students who stopped responding had the lowest average performance compared to the other two groups (35% of items correct compared to 44–45% of items correct).

Exhibit B.6: Student Response Type Classifications for Science—Grade 4

Student Response Group	Percent of Students	Time of Last Response (minutes)	Percent of Items Correct	Percent of Items Not Reached
Regular eTIMSS Science				
1. Reached All Items	93%	25.5	53%	0%
2. Ran Out of Time	3%	35.8	43%	18%
3. Stopped Responding	4%	31.8	40%	20%
Total		**25.9**	**52%**	**1%**
PSI Science				
1. Reached All Items	76%	28.1	45%	0%
2. Ran Out of Time	11%	35.9	44%	19%
3. Stopped Responding	13%	30.6	35%	22%
Total		**29.2**	**44%**	**5%**

All statistics were computed at the student level by country, then averaged across countries.
Because of rounding some results may appear inconsistent.

The results for eighth grade mathematics in Exhibit B.7 show the majority of students who took regular eTIMSS reaching all items (94%) and relatively fewer PSI students (83%) doing so. Similar to the mathematics results at the fourth grade, more eighth grade PSI students stopped responding than ran out of time (14% vs. 3%). PSI students who reached all items gave their last response at 34 minutes with 11 minutes remaining, which is similar to the regular eTIMSS students who finished at 33.6 minutes, on average. In comparison, PSI students who stopped responding last interacted with an item two minutes earlier, at 32.1 minutes with 13.9 minutes remaining, on average.

Exhibit B.7: Student Response Type Classifications for Mathematics—Grade 8

Student Response Group	Percent of Students	Time of Last Response (minutes)	Percent of Items Correct	Percent of Items Not Reached
Regular eTIMSS Mathematics				
1. Reached All Items	94%	33.6	43%	0%
2. Ran Out of Time	2%	44.9	35%	14%
3. Stopped Responding	4%	35.7	32%	16%
Total		**33.9**	**42%**	**1%**
PSI Mathematics				
1. Reached All Items	83%	34.0	29%	0%
2. Ran Out of Time	3%	44.9	32%	15%
3. Stopped Responding	14%	32.1	25%	14%
Total		**34.1**	**29%**	**3%**

All statistics were computed at the student level by country, then averaged across countries.
Because of rounding some results may appear inconsistent.

The results for eighth grade science (Exhibit B.8) were similar between regular eTIMSS students and PSI students, with the majority of students reaching all items in both groups (97% and 94%, respectively). Among PSI students, on average, the 4 percent who stopped responding had the lowest performance on the PSI tasks, with students who reached all items and ran out of time answering at least 10 percent more items correct than students who stopped responding (41–44% of items correct vs. 30% of items correct).

Exhibit B.8: Student Response Type Classifications for Science—Grade 8

Student Response Group	Percent of Students	Time of Last Response (minutes)	Percent of Items Correct	Percent of Items Not Reached
Regular eTIMSS Science				
1. Reached All Items	97%	31.1	47%	0%
2. Ran Out of Time	1%	–	–	–
3. Stopped Responding	2%	30.4	31%	17%
Total		**31.2**	**47%**	**0%**
PSI Science				
1. Reached All Items	94%	32.5	44%	0%
2. Ran Out of Time	3%	44.9	41%	15%
3. Stopped Responding	4%	31.8	30%	18%
Total		**32.7**	**44%**	**1%**

All statistics were computed at the student level by country, then averaged across countries.
A dash (–) indicates comparable data not available. Because of rounding some results may appear inconsistent.

Notes

[1] Soland, J., Kuhfield, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-scale Assessments in Education, 9*(8). https://doi.org/10.1186/s40536-021-00100-w

[2] Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement, 80*(3), 522–547. https://journals.sagepub.com/doi/full/10.1177/0013164419878241

APPENDIX C

Automated Scoring with Neural Network Modeling

Introduction

The move to computer-based assessment of student achievement allows the implementation of more complex and innovative item types that capture responses and process indicators with greater nuance. The TIMSS 2019 PSI items were developed to assess students in integrated scenarios to better measure higher-order mathematics and science skills. While advanced constructed response items promise to improve measurement, they also typically require scoring by human raters, which can be costly. Some assessments have turned to automated scoring to reduce the workload associated with human scorers, including TIMSS for short open-ended responses that can be reliably machine scored. For example, the automated scoring of items that use number pad input has improved the efficiency of the data analysis process. It is desirable to extend automated scoring beyond short number-based responses to other constructed response types to continuously improve scoring reliability while reducing human rater workload and costs.

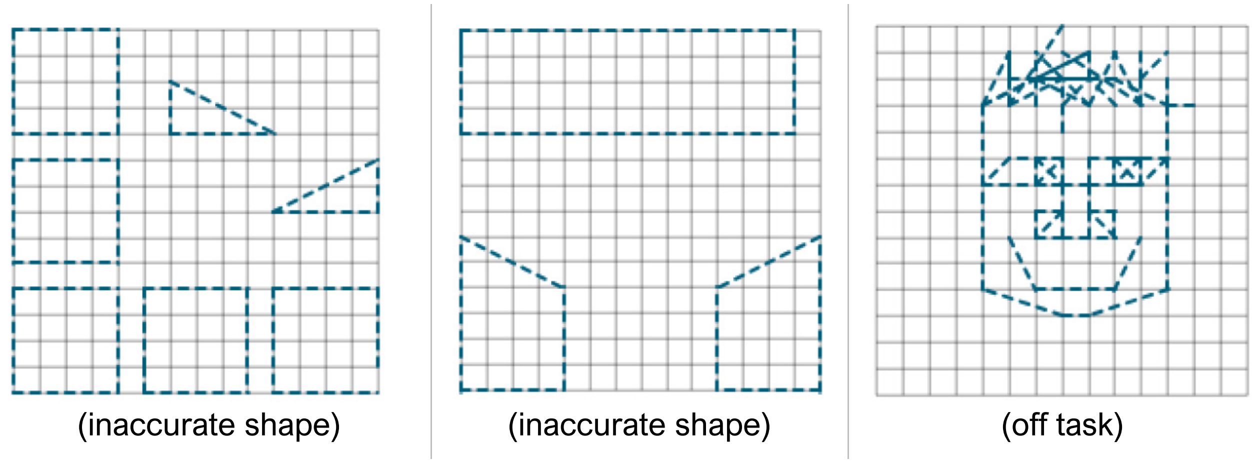
One complex item type that would benefit from automated scoring is graphical constructed response, which requires students to produce images or graphs. Developments in machine learning over the past decade have enabled algorithmic image classification, but this has yet to be utilized by any large-scale assessments. There is potential that implementing automated scoring of image-based responses improves scoring accuracy and comparability across countries by supplementing human ratings using machine approaches.

Automated Scoring of Response Images from PSI Item MQ12B03

To examine whether automated scoring could be implemented in a large-scale assessment, the TIMSS & PIRLS International Study Center conducted a study on response images from the *Building* PSI Task, Screen 4 – Constructing the Walls. This item asked students to draw a shed's back and side walls on a grid according to given specifications. Students received full credit for drawing the back of the shed and its sides correctly and partial credit if only the back wall was correct; all other responses were considered incorrect. On average, about 26 percent of students received full credit on this item by drawing all three shapes correctly; about 11 percent received partial credit. See Chapter 3 on the *Building* PSI task.

Most incorrect response images fell into three categories: blank, inaccurate shapes, and off task. Blank images were those where students had left the grid empty. For TIMSS 2019, these blank responses were automatically coded as "omitted" and were not seen by human scorers. Most of the response images contained inaccurate shapes where students had drawn shapes but did not have the correct dimensions to receive partial or full credit. Finally, off task images were those where students drew response images that were very different from the shed's walls. Exhibit C.1 displays examples of incorrect responses.

Exhibit C.1: Examples of Incorrect Responses



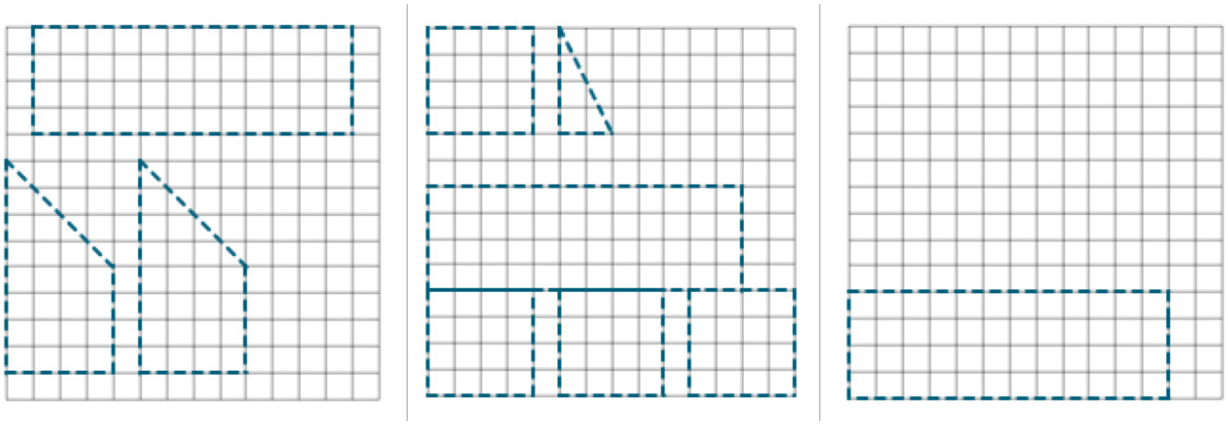
(inaccurate shape)

(inaccurate shape)

(off task)

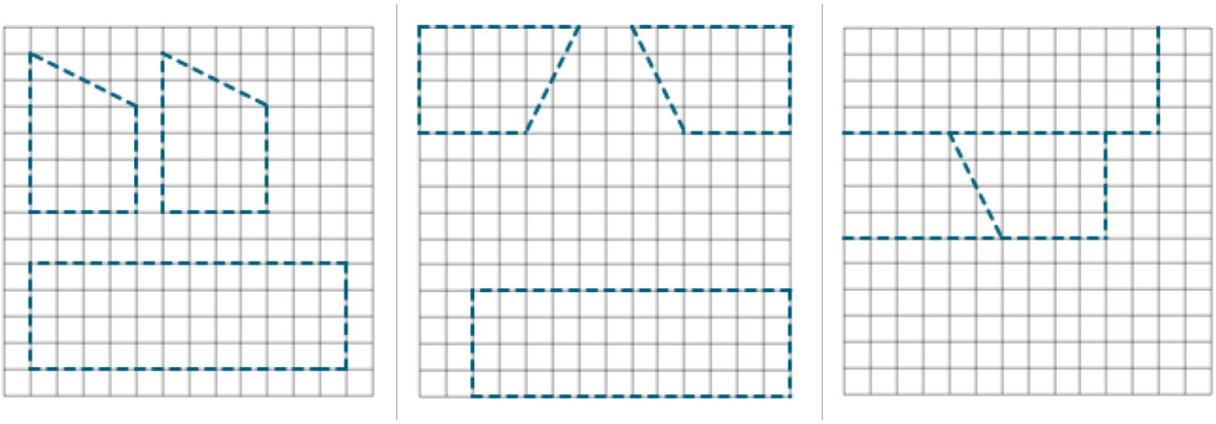
Response images given partial credit were not as diverse as incorrect responses, but they still varied from having only one shape (the back wall) to having multiple shapes. Exhibit C.2 exemplifies the variety of partial credit responses.

Exhibit C.2: Examples of Partial Credit Responses



Response images given full credit tended to be more uniform than the other two score categories, but the orientation of the shapes differed between students. While the typical response had the shapes separated in the grid, some students attempted to conserve space on the board by having the shapes be as close together as possible. Also, response images wherein the grid's border makes up one or more sides of the shape were still given full credit if the dimensions were correct. Exhibit C.3 includes examples of response images that received full credit.

Exhibit C.3: Examples of Full Credit Responses



The diversity of the response images made this item an optimal candidate for examining whether automated scoring could be applied to complex graphical responses used in TIMSS assessment items. One challenge the responses present is that the back wall and sides of the shed can be in any orientation and still receive full credit, as long as the measurements are correct. Also, the algorithm must learn that blank, inaccurate, and off task response images all have the same classification (incorrect). Finally, the two-point nature of the item could present a challenge to the learning algorithm because it would have to distinguish between three categories instead of just two.

Assignment and Pre-Processing

For the study, 14,737 response images were extracted from the data of 22 participating countries, as well as 3 benchmarking participants with unique samples. Most benchmarking participants use the same student samples as their corresponding country participant, except for Moscow City, Russian Federation, Quebec, Canada, and Ontario, Canada. The response images and associated scores produced by human raters were analyzed without any identifying information.

The evaluation of image classification followed best practices customarily applied in machine learning: The response images were assigned to two samples—10,238 were used for the first (training) sample and 4,499 were used for the second (validation) sample. The automated scoring system would be trained to classify response images on the training sample and then test the accuracy of its classifications on the validation sample. This was done on a country-by-country basis, where roughly 70 percent of response images in each score category were randomly assigned to the training sample. The remaining response images were assigned to the validation sample (see Exhibit C.4). After assignment, the response images were converted to grayscale and had their contrast enhanced to distinguish the drawn lines from the grid.

Exhibit C.4: Response Image Sample Sizes

Country	Number of Response Images		
	Training Sample	Validation Sample	Total
Chile	324	141	465
Chinese Taipei	442	161	603
England	274	124	398
Finland	373	171	544
France	304	146	450
Georgia	265	124	389
Hong Kong SAR	283	132	415
Hungary	405	173	578
Israel	312	149	461
Italy	297	156	453
Korea, Rep. of	261	115	376
Lithuania	317	151	468
Malaysia	669	289	958
Norway (9)	317	150	467
Portugal	305	117	422
Qatar	315	141	456
Russian Federation	336	138	474
Singapore	436	209	645
Sweden	323	143	466
Turkey	348	142	490
United Arab Emirates	1,831	771	2,602
United States	671	283	954
Benchmarking Participants			
Moscow City, Russian Fed.	314	157	471
Quebec, Canada	233	98	331
Ontario, Canada	283	118	401
TOTAL	**10,238**	**4,499**	**14,737**

Neural Network Modeling and Results

The automated scoring process was conducted using machine learning with artificial neural networks (ANNs). This approach was selected because ANNs are flexible and are known to achieve high accuracy for image classification tasks. Additionally, ANNs are trained to classify images with little user interaction; instead, they automatically learn to identify specific features that are correlated with certain responses. For example, an ANN model may learn that lines at a 45-degree angle of a certain length are associated with full credit responses, while images with no lines are associated with incorrect responses. This study used a particular type of ANN known as a feedforward neural network (FFN). FFNs have a simpler structure than other, more complex neural networks, and move the input information layer by layer in only one direction, which reduces processing power and improves speed.[1] More complex neural networks may include memory or feedback loops, or other structures, and are often used for sequence data or data that comes in more than two dimensions.[2,3]

The training of the FFNs and validation of the automated scoring approach was conducted in RStudio using the "Keras" package.[4,5] The automated scoring models that were examined utilized varying numbers of processing layers, iterations of training, and optimizers (which maximize feature identification based upon different mathematical formulas—"Optimizers"[6]). In total, roughly 50 models were created and their accuracies compared to identify which model had the highest number of correct classifications.

The "best" model had a three-layer structure and underwent 300 iterations of training. It was also compiled using the "nadam" optimizer.[7] This model had an accuracy of 90.35 percent and correctly classified 4,065 of the 4,499 response images in the validation sample.

This model had the most correct classifications for response images given full credit, with 94.74 percent accuracy. It correctly classified incorrect response images with 89.56 percent accuracy. However, the model only had 66.18 percent accuracy for partial credit response images. While this is noticeably lower, almost every model created in the process had the highest misclassifications for the partially correct score category. On a country-by-country basis, the models' accuracy ranged from 83.89 to 96.94 percent. Variation in accuracies across countries is expected, and there were no discerning patterns of image misclassifications. On average, 17 response images were misclassified by the model per country.

Most of the images misclassified as incorrect by the best model were rectangles with no other shapes on the grid. It appears that the model associated response images that included both rectangles and other shapes with partial credit. Response images with rectangles that were too large or too small were sometimes misclassified as partial credit, while some response images with the sides of the shed correct or those with shapes similar to the sides of the shed were misclassified as full credit.

Implications for Future TIMSS Cycles

The most accurate automated scoring neural network from the study is comparable to the accuracy human scorers achieve. Additional explorations with other types of approaches may further improve accuracy. The benefit of using automated scoring with ANNs is that it is relatively fast and cost-effective. The *Building* PSI item used in the study did not have a second independent human rater for assessing the human scoring, thus the machine scoring functioned like a second human rater and showed a high level of agreement.

Future TIMSS cycles may utilize automated scoring with ANNs in place of a second human rater. Any disagreements between the primary human raters and the machine could be reviewed and resolved by an additional expert rater at the TIMSS & PIRLS International Study Center. This expert rater would only need to review a fraction of the responses that full-time raters would.

To utilize ANNs operationally for automated scoring, models could be trained on response images from past cycles or, for new items, on response images from the field test data collections. Then the most accurate model would be applied to the response images collected during the actual studies. It should be noted that while response sample sizes will be smaller from a field test than actual data collection, images can be transformed (e.g., rotated, flipped, and cropped) to increase the training data's sample size.

Limitations

One limitation of using automated scoring with ANNs is that the training data must be accurate. Early in the study, it was found that some response images were scored incorrectly by human raters. Additionally, there were some scoring inconsistencies across countries, particularly for response images that had the shed's walls with the correct dimensions but also included extraneous lines. These inconsistencies could be due to some raters giving students the benefit of the doubt more than others. Incorrect and inconsistently scored images were removed from the training and testing samples before the study began ANN modeling with all of the countries' response images.

For future TIMSS cycles, training data must be carefully assessed so that incorrect and inconsistent response images are identified and re-scored. This process will likely require at least one additional human rater. However, the workload and associated cost would still be less than full sample double scoring by human raters to review every response image.

Notes

[1] Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems, 39*(1), 43–62. https://doi.org/10.1016/S0169-7439(97)00061-0

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[3] von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*, 847–857. https://doi.org/10.1007/s11336-018-9608-y

[4] Chollet, F. (2017). *Deep Learning with Python*. Shelter Island, NY: Manning Publications.

[5] Kalinowski, T., Allaire, J., & Chollet, F. (2021). *R Interface to Keras*. Keras. https://keras.rstudio.com/index.html

[6] Keras. (2021). *Optimizers*. Keras API Reference. https://keras.io/api/optimizers/

[7] Dozat, T. (2015). *Incorporating Nesterov momentum into Adam*. http://cs229.stanford.edu/proj2015/054_report.pdf

BOSTON
COLLEGE

timss.bc.edu

IEA

© IEA, 2021
International Association
for the Evaluation of
Educational Achievement

ISBN-978-1-889938-58-5