2013-07-10

The United States Growth over 16 Years of Student Correct Responses on the TIMSS: Are We Really That Far Behind?

Jacob Michael Zonts
Brigham Young University - Provo

The United States Growth over 16 Years of Student Correct Responses on the TIMSS:

Are We Really That Far Behind?

Jacob M. Zonts

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Nancy Wentworth, Chair
Roni Jo Draper
Janet Young

Department of Teacher Education

Brigham Young University

July 2013

ABSTRACT

The United States Growth over 16 Years of Student Correct Responses on the TIMSS: Are We Really That Far Behind?

Jacob M. Zonts
Department of Teacher Education, BYU
Master of Arts

National rank on international assessments, as measured in Trends in International Mathematics and Science Study (TIMSS), gives a limited view of the data presented. This study used average scale score data from the TIMSS (1995, 1999, 2003, 2007, 2011) that were then disaggregated based on content domains (i.e., number, algebra, measurement, geometry, data, earth science, life science, physical science, biology, physics, and chemistry). These data were graphed to show the growth of the U.S. national average scale scores in comparison to three top scoring countries (i.e., Hong Kong, Japan, and Singapore), and three other post-industrial countries similar to the U.S. (i.e., England, Italy, and Australia) It was found that the eastern nations outperformed the western nations on science math question for the fourth and eighth grade. The gap between eastern and western nations grew from the fourth to eighth grades. For fourth- and eighth-grade science content domains, Singapore outperformed all other nations except in earth science where all nations were evenly matched. Additionally, percent correct statistics from the *2011 TIMSS Released Items* were disaggregated based on subject (i.e., science and mathematics) and cognitive domain (i.e., knowing, applying, and reasoning). The released item scores, based on cognitive domain, were then averaged and the U.S. averages were compared with the averages of the above mentioned nations, using a series of *t*-tests. Singapore scored significantly higher in all categories except fourth-grade science cognitive domains knowing and applying. Hong Kong scored significantly higher in fourth- and eighth-grade mathematics cognitive domains knowing and applying and eighth-grade mathematics cognitive domain reasoning. Japan scored significantly higher in eighth-grade mathematics cognitive domains applying and reasoning as well as science cognitive domain applying. These findings suggest that the U.S. is lagging behind in some content domains and cognitive domains, but not all. The current study informs teachers, administrators, and policy makers of the specifics areas the U.S. needs improvement.

Keywords: [TIMSS, 2011, rank, Singapore, U.S., international, assessment, content domain, cognitive domain, *t*-test, trend line]

Table Of Contents

List Of Tables

List Of Figures

Chapter 1

Introduction

Public education in the U.S. is under fire by education reform groups (e.g., National Commission on Excellence in Education, 1985), news organizations (e.g., Armario, 2010; Rich, 2012; Gillespie, 2013), and politicians (e.g., U.S. Department of Education, 2010; Nagesh, G. 2010.) These documents, news articles, and speeches frequently quote the poor rank of U.S. students on international tests, such as the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA). They use this rank to justify areas they think schools need to focus, such as accelerating achievement in secondary school and closing the achievement gaps among black and Hispanic students (Carnoy & Rothstein, 2013). The rank of a nation is often the first datum that is released after an international test and will often be the first piece of evidence reported.

International Tests and the Problem of Relative Rank

The relative rank of a nation can give policymakers and the general public a quick and simple view of a large set of data. It is often very easy for researchers to report rank on tests because of the simplicity of the data analysis being reported. In December of 2012 the International Association for the Evaluation and Educational Achievement (IEA) released a very small portion of the results of the administration of the 2011 TIMSS. The release only included the national average results from the test. This information showed the relative rank of U.S. students on the test as well as the relative rank of all other participating countries. Similar to previous years releases, the IEA released the relative rank of countries far in advance of the international database. The published database allows researchers the ability to disaggregate test scores based on a multitude of student and test item characteristics. Quick-view statistics, such as rank, that policymakers and news organizations latch onto so quickly, do not include statistics that have used disaggregated test scores. The rank is based solely on the overall achievement of all participating students.

The inability of political pundits and politicians to make reasonable conclusions based on rank does not stop them from making statements based entirely on those conclusions. Following the release of the national average results from the 2011 administration of the TIMSS, U.S. Secretary of Education, Arne Duncan, in a press release (2012) stated that "these new international comparisons underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps" and then stated that these results are "unacceptable" (para. 5). Using rank as the crux of his argument, his statements are not as reliable as they would have been if he had waited for research that used disaggregated test results.

Alan Smithers of the Centre for Education and Employment Research of the University of Buckingham states, "TIMSS rankings... are based on raw scores which can make it appear that a country has done better when it has not. A difference of a point or two in the mean score can have a disproportionate effect on the ranking. But raw scores are subject to random variation, and an apparent difference may not be real" (2013, p. 33). The TIMSS, over the past five test cycles (i.e., 1995, 1999, 2003, 2007, and 2011), has given a plethora of data that need to be disaggregated. Relative rank touches on only a very small portion of the complex data that are collected when an international assessment is implemented.

A wide variety of comparisons of international test results in education have been made since the beginning of international tests in education. Curtis C. Mcknight et al. (1987), researchers who conducted the Second International Mathematics Study (SIMS), published a report that made comparisons between the nations that participated in the study based on the content of the intended mathematics curriculum (i.e., arithmetic, algebra, geometry, measurement, and calculus). Dividing the test scores based on a variety of subjects, McKnight was able to ascertain a clearer picture of the data that were in the SIMS. Many other comparisons where made throughout the years. These include comparisons of curriculum (Burstein,1993; Pelgrum, Eggen, & Plomp, 1986), comparisons between content areas (Beaton, et al., 1996; Beaton, et al., 1997) and comparisons based on gender (Neushmidt, Barth, & Hatedt, 2008), to name a few. Comparisons between nations over time as well as comparisons based on cognitive domains are lacking in all of these studies.

Until recently, a comparison of nations over time has been difficult. The difficulties arise for several reasons. For instance, there was a large gap in time between international tests before 1995, making comparison over time unreliable. Other problems that have made comparisons difficutlt include the fact that very few nations have taken all or even most of the international tests. Further, differences in how the tests measured achievement has changed over time as well. Until the most recent iterations of the TIMSS, international tests evolved in the way they measured student success. The most recent iterations of the TIMSS specifically, 2007 and 2011, have changed the way student success is measured by organizing test results based on cognitive domains. Student success on questions of cognitive domain can now be measured. Furthermore, student success can now be compared through the years based on content domain. This is possible because of the standardization content domains over the past decade. Content domains now have a one-to-one correlation across time.

There are very few studies that disaggregate the TIMSS test data into content and cognitive domains and then compare nations over time. The TIMMS lends itself well to a study of this type in that the TIMSS has been implemented five times over the past eighteen years. IEA has released massive databases for each of the five tests, which have allowed researchers to test hypothesis and make conclusions based on the data.

In order for the public and policymakers to make educated decisions about problems with education that face the nation, they must have findings specifically geared to those problems. A relative rank on an international test only tells a very small story within a large collection of data. Conclusions based on ranking alone are "oversimplified, frequently exaggerated, and misleading. They ignore the complexity of test results and may lead policymakers to pursue inappropriate and even harmful reforms" (Carnoy & Rothstien, 2013, para. 4). One area of study that has not been pursued is a comparison of average national scores based on content domain over time. Also, comparisons based on cognitive domain have not been performed. If these two gaps in the research are filled, the public and policymakers will have a better understanding of the problems facing education and be better able to come up with solutions to those problems.

Research Questions

The purpose of this study was to examine the effectiveness of national ranking on international tests (i.e., TIMSS 1995, 1999, 2003, 2007, 2011) through a comparison of ranking, first with a comparison of average international test scores of the TIMSS (i.e., 1995, 1999, 2003, 2007, 2011) disaggregated into results based on content domains, and next, with a correlation between percent student correct responses of the TIMSS (2011) between nations based on cognitive domain.

The purpose of this study was to bring forward evidence that goes beyond the use of rank alone as a measure of a nation's academic achievement in science and mathematics. This study includes data from the 2011 published TIMSS database, which have not been used in any published study to date. Therefore, I put forth these questions and have sought to find evidence in order fulfill the aforementioned purpose:

1. What are the differences in growth over time between the U.S., three top scoring countries (i.e., Hong Kong, Japan, and Singapore), and three other post-industrial countries similar to the U.S. (i.e., England, Italy, and Australia) on average correct responses of questions within the same content domains of the TIMSS for the years 1995, 1999, 2003, 2007, and 2011?
2. Based on question of similar cognitive difficulty (i.e., knowing, applying, reasoning) from the 2011 TIMSS, was the U.S. significantly different in comparison to three top scoring countries (i.e., Hong Kong, Japan, and Singapore) and three other post-industrial countries similar to the U.S. (i.e., England, Italy, and Australia) on percent student correct responses?

In order to answer question 1, I used the TIMSS databases, for the years, 1995, 1999, 2003, 2007, and 2011. Within the published databases, IEA has given average scale scores for all of the nations that participated in each year based on content domains. The content domains used in the TIMSS are life science, physical science, and earth science for fourth-grade students in science; biology, physics, chemistry, and earth science for eighth-grade students in science; number, geometry, and data for fourth-grade students in mathematics; and number, geometry, data, and algebra for eighth-grade students in mathematics. Using the average scale scores for each of these grades and content areas per country, I plotted the scores for each year of the TIMSS. I drew conclusions on the apparent positive or negative growth of nations over time.

In order to answer question 2, I used the *TIMSS Released Items* database for 2011. This document contains individual items released by IEA along with item statistics that contain the percent student correct response for each country per question. The statistics contain data for fourth and eighth grade with delineations of cognitive domain for each question. Using the item statistics I calculated the significant difference between correct student responses based on questions of cognitive difficulty (i.e., knowing, applying, and reasoning) of the U.S. and the other nations included in the current study. I used *t*-tests to determine significant difference between nations.

The next two chapters elaborate on the different aspects of the current study. Chapter 2 includes a discussion of the literature that has led up to and finds a need for this study, specifically the nature of the TIMSS and the research that has been conducted using its databases. Chapter 3 discusses the research methodology and design of this study. Chapter 4 presents the findings based on the data. Chapter 5 discusses the implications of the current study.

Chapter 2

Review of the Literature

National rankings on international tests have become a major component of educational reform in the U.S. This review of the literature seeks to establish a basis for my current inquiry into the effectiveness of rank of the U.S. on international tests, specifically the TIMSS. Through the history of TIMSS, researchers have delved into the databases that have been created by the TIMSS. They have disaggregated results based on demographic background (Blomeke, Suhl, Kaiser, & Dohrmann, 2012), student achievement (Tienken, 2013), content domains (Suter, 2000), and gender (Mullis, Martin, Fierros, Goldberg, & Stemler, 2000), to name a few. Within each of these separate studies, ranking was not a major portion of the data analysis. Rather researchers have used the databases to make inferences about many topics covered in the TIMSS. Only a handful of researchers have questioned the validity of overall national ranking on international tests (Bracey, 2007).

Although there are studies that look at the relative ranking on the TIMSS, there are no studies, outside of the IEA released documents, which include the 2011 database in their data set. In addition to the lack of studies that use the 2011 data, there are also few studies that have looked at the growth of average national scores on the TIMSS, and none that have included the 2011 data. In the following paragraphs I go further in depth in each of these fields to examine the gap in knowledge that exists in the body of educational research and also includes a brief history of the TIMSS.

History of the TIMSS

The TIMSS had its beginnings in the late 1950s under the name of First International Mathematics Study (FIMS) (Amrein, & Berliner, 2002). The FIMS, created by the IEA, was the first international comparative study of educational achievement of its kind. The idea of a large-scale cross-national study of mathematics began with the research of Benjamin S. Bloom who presented his ideas to his colleagues in 1958 (Husen, 1967). Although the FIMS was able to make comparisons between nations, comparisons over time would not be possible until the SIMS. SIMS was being discussed by the middle of the 1970s. Beginning in 1976, IEA researchers started to create the SIMS and subsequently collected data from 1980-82.

Implemented in 1995, TIMSS was the first to test both mathematics and science in the same study. Not only was science included in this study, but both science and mathematics received a large increase in assessment questions. The study was also able to include three levels of the education system, grades three and four (the end of primary schooling), grades seven and eight (middle school) and twelfth grade (or final grade). The twelfth-grade students were not tested every year of the TIMSS and were at times tested years apart from the other grades. Over time funding for TIMSS increased. This resulted in an increase in the number of assessment questions on the tests. Because of the increased number of assessment questions, the validity of disaggregated results increased. Additional questions being used on the TIMSS meant more data per content domain. With additional questions for each content domain being tested, researchers are now able to compare international scores based on content domains.

Following the 1999 administration of the TIMSS, IEA decided to perform the study every four years. Subsequently, the name of the study was changed from *Third International Mathematics and Science Study* to *Trends in International Mathematics and Science Study*. The TIMSS has now completed cycles for the years 1995, 1999, 2003, 2007, and 2011, with final preparations being made for the 2015 administration. Thus, researchers can now study the national trends over the course of five studies. For the purpose of the current study I will only compare nations based on cognitive and content domains.

Content and Cognitive Domains

The mathematics and science components of TIMSS 1995, 1999, 2003, 2007 and 2011 are organized into two dimensions: content domains and cognitive domains. In this section a summary of these domains is given (see Table 1). Within each domain, there exist a number of components that make up the domain s a brief definition of each component is also included. These definitions are based on the TIMSS frameworks that were published along with the TIMSS databases for each iteration of the test. In addition to a summary, this section includes a sampling of relevant research studies that have used the content and cognitive domains as part of their disaggregation of the extremely complex data of the TIMSS databases.

The first dimension of TIMSS is the content domain. Each of the TIMSS (i.e., 1995, 1999, 2003, 2007, and 2011), reports on the average score for each participating country based on content domain. The content domain components are: life science, physical science, and earth science for fourth-grade students in science; biology, physics, chemistry, and earth science for eighth-grade students in science; number, geometry, and data for fourth-grade students in mathematics; and number, geometry, data, and algebra for eighth-grade students in mathematics. Because these domains are an integral to the current study, common definitions are established here. Table 1 contains the definitions established by IEA during the design of the TIMSS.

Cognitive domain components consist of knowing facts, procedures, and concepts; applying knowledge and understanding; and reasoning. Because these domains are integral to the current study, common definitions are delineated here, as established by IEA during the design of the TIMSS.

Table 1
Definitions of Content Domains Established by IEA

Science	Mathematics
Life Science. The life science domain includes understandings of the characteristics and life processes of living things, the relationship between them, and their interactions with the environment (Martin, et. al., 2008).	Number. The number domain consists of whole numbers, fractions and decimals, integers, ratio, proportion, and percent (Mullis et. al., 2003).
Biology. The biology domain includes students' understandings of the structure, life processes, diversity, and interdependence of living organisms (Martin et al., 2008).	Geometry. The geometry domain includes understanding "lines and angles, two- and three-dimensional shapes, congruence and similarity, locations and special relationships, symmetry and transformation" (Mullis et al., 2003, p. 18).
Physical Science. The physical science domain includes concepts related to matter and energy, and covers topics in the areas of both chemistry and physics for the fourth grade (Martin, et al., 2008).	Data. The data domain covers "data collection and organization, data representation, data interpretation, and uncertainty and probability" (Mullis et al., 2003, p. 21).
Physics. In the physics domain, students' understanding of concepts related to energy and physical processes were assessed based on physical states and changes in matter, energy transformation, heat, and temperature, light, sound, electricity and magnetism, forces and motion (Martin et al., 2008).	Algebra. The algebra domain consists of "patterns and relationships among quantities, using algebraic symbols to represent mathematical situations, and developing fluency in producing equivalent expressions and solving linear equations" (Mullis et al., 2003, p. 14).
Chemistry. In the chemistry domain, students were assessed on their understanding of concepts related to classification and composition of matter, properties of matter, and chemical change (Martin et al., 2008).	
Earth Science. The earth science domain is concerned with the study of Earth and its place in the solar system, which include earth's structure, and earth's processes (Martin et al., 2008).	

As Mullis and associates (2005) point out,

> Knowing facts, procedures, and concepts, covers what the student needs to know, … applying knowledge and conceptual understanding, focuses on the ability of the student to apply what he or she knows to solve routine problems or answer questions..., reasoning, goes beyond the solution of routine problems to encompass unfamiliar situations, complex contexts, and multi-step problems (p. 7.)

Several studies have been performed that classified test items into content knowledge, process and skill attributes. In one such study, Tatsuoka, Corter and Tatsuoka (2004) examined the mathematics items of TIMSS 1999 based on 23 specific content knowledge and processing domain. Mean mastery levels for each domain were compared for 20 selected countries. They found that the U.S. students were strong in some content domains but weak in others, notably geometry. Chen, Gorin, Thompson, & Tatsuoka (2008) used content domains to classify achievements made by Taiwan students on the TIMSS 1999. Chen et al.(2008) found that Taiwanese students consistently showed strengths and weaknesses in certain content domains. These two studies give us a better understanding of where certain countries stand with regards to content knowledge as well as process and skill attributes but they lack the ability to see trends over time because they used data from a single year of the TIMSS.

Although many studies have focused on content domains as a means to understand the level of proficiency of a national education system, very few studies compare national scores based on cognitive domain other than the TIMSS Reports. An example of a study which focused on cognitive domains is the study performed by Toker (2010) in which he used released items from the 2007 TIMSS to administer a test to current eighth grade students in order to validate the national score in comparison to the participants of his study. The study found there was no significant difference between the different cognitive domains, and as a result certain deficiencies where found in the education system of Turkey. Although this study used cognitive domains as a way to disaggregate the items, the study did not use the TIMSS data as the main source of data but instead used scores from a repeat of released test questions.

The TIMSS reports contain the disaggregated scores for each of the cognitive domains for TIMSS 2003 as well as the scores for individual test items. Although the TIMSS reports contain the score for individual items with the cognitive domain stated, the authors of the reports make no effort to make comparisons within and across countries. I will be able to compare the percent student correct responses for each cognitive domain with the relative national rank on the international test. This would give another layer of evidence as to discern whether overall rank is suitable for policymakers to make changes in policy.

These example studies and reports indicate that the TIMSS data have been used to uncover meaning on the content and cognitive levels. These studies and reports have added to our understanding of the problems that face our country with regards to education, but overall they each look at a snapshot of student achievement instead of looking at the growth made over time.

Many studies have been performed using the data found in the TIMSS 1995, 1999, 2003, 2007 databases. Almost no studies have been done using the 2011 database. With the recent release of the 2011 database, many of the studies that have been mentioned could be performed again with the inclusion of the 2011 data. Inclusion of the 2011 data is a major justification for the need of this study. In addition to the increase in data because of the 2011 data release, this study seeks to expand the understanding of the usefulness of relative national ranking on international tests.

In lieu of overall ranking on an international test such as TIMSS, researchers could focus on a number of other attributes that would give them an understanding of the relative standing of a certain nation compared to other nations. Researchers could focus on comparisons of cognitive domain, affective domain, background domain, content knowledge, curriculum, teacher effectiveness, home life, socioeconomic status, race, gender and many other attributes. Data for all of these domains are provided in the TIMSS databases. Each of these attributes has a specific research goal in mind when they are being studied. For this study I could have used any number of these attributes but has chosen to focus on the comparison of average score based on content and cognitive domains. These domains have been studied less than some of the other domains mentioned. They can also be used to gain another type of ranking that can be compared easily to the original overall ranking. Also, the content domains are easily observed over time and can be graphed to show the growth that nations have made on their average score based on content domains.

There are many ways in which I could expand the meaning of the national rank of the many countries involved in the TIMSS, as can be seen by the prior research that has delved deeper into the TIMSS data, but I chose to use average score based on content domains as well as cognitive domains as the means to accomplish the original purpose of this study.

As stated in chapter 1, the purpose of this study is to examine the effectiveness of national ranking on international tests (i.e., TIMSS 1995, 1999, 2003, 2007, 2011). Previous findings of recent research lead me to the conclusion that ranking is in question (Carnoy &

Rothstein, 2013, para. 4) and that "international test comparisons are oversimplified, frequently exaggerated, and misleading. They ignore the complexity of test results and may lead policy-makers to pursue inappropriate and even harmful reforms" (Carnoy & Rothstein, 2013, para. 5). Through additional data analysis, using data from every year of the TIMSS including 2011, I compared the relative rank of nations with the disaggregated average scores based on content and the percent student correct statistics for the TIMSS 2011 released items disaggregated based on cognitive domain. The details of this process will be discussed in the following chapter.

Chapter 3

Methods

Research on TIMSS has not included more current data from the 2011 test. Therefore, inclusion of the 2011 data is a major justification of the current study.

For this study I could have used a variety of attributes to compare nations, such as cognitive domain, affective domain, background domain, content knowledge, curriculum, teacher effectiveness, home life, socioeconomic status, race, gender. I chose to focus on the comparison of average score based on content and cognitive domains. These domains have been studied less than some of the other attributes mentioned. They can also be used to gain another type of ranking that can be compared easily to the original overall ranking. Further, the content domains have one-to-one correlation over time, which allow the data to be graphed to show the growth that nations have made on their average score.

As stated in chapter 1, the purpose of the current study was to examine the effectiveness of national ranking on international tests (i.e., TIMSS 1995, 1999, 2003, 2007, 2011). Through additional data analysis, using data from every year of the TIMSS including 2011, I compared the relative rank of nations with the disaggregated average scores based on content and the percent correct statistics for the TIMSS 2011 released items disaggregated based on cognitive domain. The details of this process are discussed in the following sections.

Design

The current study used two quantitative research methodologies. Research question 1: growth of average scale scores over time used a methodology which includes the formation of trend lines of multiple nations. The trend lines were then analyzed using a series of questions that deal with apparent growth over time. These questions are discussed further in subsequent sections of this chapter. Research question 2: comparison between the U.S. and selected countries across cognitive domains used a methodology based on statistical analysis, specifically tests for significant difference. After the series of *t*-tests were performed, I was able to specify more narrowly the areas in which the U.S. is lagging. These two methodologies fit well with the data I chose to use, which were the average scale scores of nations based on content domains and the percent correct statistics for the TIMSS 2011 released items disaggregated based on cognitive domain.

Participants

Because of the large number of nations who have participated over the course of TIMSS, the current study's analysis were restricted to the test results of fourth- and eighth-grade students in mathematics and science in the U.S., in three top-scoring countries, and in three other post-industrial countries similar to the U.S. These countries are a sampling of countries with which the U.S. is most often compared. Many other nations that fit these criteria had to be eliminated because they had not participated in all of the TIMSS tests for the years in which the current study pulls data. One such nation that is often referred to as a nation that leads in international tests is Finland (Robelen, & Gewertz, 2013; Robelen, 2013), which only participated in the 1995 and the 2007 TIMSS. Therefore, Finland was not included in the current study.

One exception to the criteria was made with the inclusion of Italy in the current study. Italy did not have fourth-grade students participate in the 1995 TIMSS. I included Italy because it participated in all other years and grades of the TIMSS. The 1995 TIMSS included the fewest number of countries. Because of the low international participation, only five countries besides the U.S. were found to have participated in all other iterations of the TIMSS. I chose to include Italy because it met all other criteria with the exception of participation in the fourth-grade 1995 TIMSS.

Students in the fourth and eighth grades were chosen to participate in the TIMSS and subsequently provided the statistics for the current study. A sampling of students in each nation was chosen to represent each nation as a whole. Due to the difference in population of the many nations who have participated in the TIMSS and the amount of funding each nation allocated for the study (each participating country was responsible for funding national project costs and implementing TIMSS in accordance with the international procedures established by IEA), each nation's sample size varies. In each nation the students in both the fourth and eighth grades answered questions within the subjects of mathematics and science. In addition to subject matter questions, the students also answered questions based on demographics. All of these data were compiled into databases. These databases are the sources of statistics used in the current study.

Research Question 1: Growth of Average Scale Scores Over Time

The data sources, procedures, and data analysis are very different for each research question. Therefore, the current and subsequent section are separated based on research question. For each research question, data sources, procedures, and data analysis are presented.

Data sources. For research question 1: growth of average scale scores over time, Table 2 shows the titles of database documents used in the current study. I used the published TIMSS database documents for each of the years in which TIMSS was administered (i.e., 1995, 1999, 2003, 2007, and 2011). For each year of the TIMSS, the published database documents were divided into mathematics and science. In addition to the subject matter division, the 1995 documents were divided based on grade level.

Table 2
Titles of the TIMSS Database Documents Used in the Current Study

Year	Mathematics	Science	Release Items
1995	*Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study* (Mullis et al., 1997)*	*Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study* (Martin et al., 1997)*	NA
	Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (Mullis et al., 1996)*	*Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study* (Martin et al., 1996)*	NA
1999	*TIMSS 1999 International Mathematics Report Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade* (Mullis et al., 2000)*	*TIMSS 1999 International Science Report Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade* (Martin et al., 2000)*	NA
2003	*TIMSS 2003 International Mathematics Report Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades* (Mullis et al., 2004)*	*TIMSS 2003 International Science Report Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades* (Martin et al., 2004)*	NA
2007	*TIMSS 2007 International Mathematics Report* (Mullis et al., 2008)*	TIMSS 2007 International Science Report (Martin et al., 2008)*	NA
2011	*TIMSS 2011 International Results in Mathematics* (Mullis et al., 2012)*	*TIMSS 2011 International Results in Science* (Martin et al., 2012)*	TIMSS 2011 Release Items**

Note. *published database documents that were used for research question 1. **published database documents that were used for research question two.

Procedures. For the current study I made comparisons between the U.S. and other nations based on content domains. The content domains for TIMSS include physical science, life science, and earth science for fourth-grade students in science; chemistry, biology, physics, earth science and for eighth-grade students in science; number, geometry, and data for fourth-grade students in mathematics; and number, geometry, data, and algebra for eighth-grade students in mathematics. Data within the results documents were disaggregated into average national score based on content domains for each of the five years used in the current study. This provided a single data point for each year of the TIMSS based on content domain. I chose content domain as my comparison factor because there is an alignment though the five iterations of the TIMSS. This alignment allowed me to observe the changes over time of each nation's average scale score for each of the several content domains.

The current study presents the results at the fourth and eighth grades of growth in average scale scores of nations on the TIMSS. The results presented here show the growth over time of average scale scores.

Data Analysis. Using the data points that represent the average scale score of each nation based on content domain and grade level for each year of the TIMSS, I created 18 graphs (i.e., Figures 1-18). Each nation has a data point for each year of the TIMSS on each graph, with the exception of Italy. Italy did not participate in the 1995 administration of TIMSS for the fourth grade as previously mentioned. Thus, Italy is missing the first data point for all of the graphs that represent the data from the fourth-grade administration of the TIMSS. The 18 graphs can be divided into four groups of graphs. Fourth-grade science scores (i.e., Figures 1-4). Eighth-grade science scores (i.e., Figures 5-9). Fourth-grade mathematics scores (i.e., Figures 10-13). Eighth-grade mathematics scores (i.e., Figures 14-18).

The format of each graph was standardized in order that observations could be easily made between the different content domains and grade levels tested. In order to make more reliable conclusions, the TIMSS researchers converted the results into average scale scores. The range of average scale scores is between 1 and 1000 with the international average at 500. Since the data points centered around 550 and did not go above 650 or below 450 in the current study, I chose to make the range for the y-axis between 400 and 700. The x-axis gives the year in which data were collected. Each data point was then plotted and a trend line was drawn to connect those points.

After the data points were plotted and a line was drawn to connect those points, analyses were made based on several questions:

- Are there nations that consistently outpace all other nations?
- Are there nations that consistently lag behind all other countries?
- Are there groups of nations that move between multiple rankings throughout the five iterations of the TIMSS?
- Is there an overall positive or negative growth pattern of scores over time?
- Is there a convergence of scores over time?

Answering these questions allowed me to answer the larger question of: What are the differences in growth over time between the U.S., three top scoring countries (i.e., Hong Kong, Japan, and Singapore), and three other post-industrial countries similar to the U.S. (i.e., England, Italy, and Australia) on average correct responses of questions within the same content domains of the TIMSS for the years 1995, 1999, 2003, 2007, and 2011?

Research Question 2: Comparison Between the U.S. and Selected Countries Across Cognitive Domains

Data sources. For research question 2: comparison between the U.S. and selected countries across cognitive domains, Table 1 shows the titles of database documents used in the current study. I used the published TIMSS 2011 released items. The *TIMSS 2011 Released Items* document was divided into four sections. For each subject, mathematics and science, the document had two sections. The first section included each released test item along with the descriptive details stating the cognitive domain that the question represented. The second section of the document contained the percent of student correct responses statistics for each item. The percent correct for each country was included in this section. An example of a released test item and an example of the item statistics can be found in Appendix A and Appendix B, respectively.

Procedures. The same three cognitive domains—knowing, applying, and reasoning—were used at both the fourth and eighth grades.

> Knowing refers to the student's knowledge base of mathematics facts, concepts, tools, and procedures. Applying focuses on the student's ability to apply knowledge and conceptual understanding in a problem situation. Reasoning goes beyond the solution of routine problems to encompass unfamiliar situations, complex contexts, and multi-step problems. (Mullis et al., 2012, p. 140)

For research question 2, I used data collected for the TIMSS year 2011. This is because the 2011 and 2003 administration of the TIMSS were the only years in which item statistics have been attached to the released question. The 2011 data were chosen over the 2003 data because they are the most recent data to have been released. The test items that were released were a portion of the 2011 TIMSS and would not be included in subsequent years of the TIMSS. The released test items are only a portion of total test items used in the 2011 TIMSS.

For each of the test items that have been released, there is a detailed description (see Appendix A and B). Included in the description is content domain, main topic, cognitive domain, overall percent correct with a list of countries and their respective scores, the question, and the questions answer key. For this study, I used the cognitive domain and overall percent correct sections for each released test item. This allowed me to disaggregate the scores based on cognitive domain. It also allowed me to obtain a group of scores that could then be used to find a mean score. A number of the test items contained multiple scored questions. For the current study I treated each question within a test item as an individual test item. I used all of the available test questions and the data associated with each question that is available for the 2011 TIMSS. The following are included in the *TIMSS Released Items* document are:

- 36 fourth-grade science cognitive domain knowing test items
- 37 fourth-grade science cognitive domain applying test items
- 11 fourth-grade science cognitive domain reasoning test items
- 35 eighth-grade science cognitive domain knowing test items
- 42 eighth-grade science cognitive domain applying test items
- 22 eighth-grade science cognitive domain reasoning test items
- Total of 84 fourth-grade science test items
- Total of 99 eighth-grade science questions
- 29 fourth-grade mathematics cognitive domain knowing test items
- 29 fourth-grade mathematics cognitive domain applying test items
- 15 fourth-grade mathematics cognitive domain reasoning test items

- 33 eighth-grade mathematics cognitive domain knowing test items
- 33 eighth-grade mathematics cognitive domain applying test items
- 24 eighth-grade mathematics cognitive domain reasoning test items
- Total of 73 fourth-grade mathematics test items
- Total of 90 eighth-grade mathematics questions

Analysis to address research question 2 required that data be a list of scores that could be used to determine a mean score. Appendix C contains a table of identification numbers found in the *2011 TIMSS Released Items* document for each test item used. Some identification numbers end with a letter. The letter identifies the question within an individual test item. As mentioned previously, for the current study each question within an individual test item is considered a test item. For each nation, a percent student correct statistic for each released test item is included in the *2011 TIMSS Released Items* document. Appendix B gives an example of a released item statistics table. After obtaining the percent correct statistic for each released test item of each nation, the scores were then disaggregated based on cognitive domain (i.e., knowing, applying, and reasoning) and grade level. The percent student correct statistics were then averaged, providing the necessary mean scores to perform a series of *t*-tests.

Data Analysis. With the 346 TIMSS released test item statistics disaggregated based on subject (i.e., science and mathematics), grade level (i.e., fourth and eighth), and cognitive domain (i.e., knowing, applying, and reasoning). A total of 12 lists of percent student correct responses were created for each of the nations included in the current study (i.e., fourth-grade science knowing, fourth-grade science applying, fourth-grade science reasoning, eighth-grade science knowing, eighth-grade science applying, eighth-grade science reasoning, fourth-grade mathematics knowing, fourth-grade mathematics applying, fourth-grade mathematics reasoning, eighth-grade mathematics knowing, eighth-grade mathematics applying, and eighth-grade mathematics reasoning). For each group of lists of percent student correct responses based on nation, a series of *t*-tests were performed in order to test for significant between the U.S. and all other nations included in the current study. This was repeated for each of the 12 groups of released test items.

Results of the 72 *t*-tests were examined in order to answer the question: Was the U.S. significantly different from Singapore, Japan, Hong Kong, Italy, England, and Australia based on questions of similar cognitive ability?

Limitations

Although this study sought to include the latest year of the TIMSS, it did not seek to use all of the available data. Specifically, it did not use the majority of the nations that participated in the test. This limitation was necessary in order to conduct the research within the time-frame allotted for the current study. Also, by limiting the nations who participate in this study, I was able to make comparisons of nations that are similar to the U.S. By choosing nations that are in a similar first world situation, I was able to present a clearer picture of the needs that the U.S. may have with regards to educational reform. Countries that are not in similar socio-economic situations and that have vastly different cultures would not have given an accurate comparison.

Other limitations include the small sampling of test questions from the 2011 published TIMSS database. It was unavoidable that I was unable to receive all of the test questions. IEA only releases test items that they do not intend to use in future iteration of the TIMSS. Another limitation to this study is the fact that the demographics of each of the participating nations were different as well as the cultural background of the students from the different nations. These differences may have caused some of the test questions to be answered incorrectly.

Chapter 4

Findings

The TIMSS mathematics and science assessments are organized around two dimensions: a content dimension specifying the content domains or subject matter being assessed, and a cognitive dimension specifying the thinking process by which the students are likely to use when they are answering a question. The first section of this chapter will present the findings for research question 1: growth of average scale scores over time. The second section of this chapter will present the findings for research question 2: comparison between the U.S. and selected countries across cognitive domains. Following these two sections will be a summary of findings.

Research Question 1: Growth of Average Scale Scores Over Time

Fourth-grade science scores. This section contains the figures that present the fourth-grade science scores. Figure 1 presents the overall fourth-grade science scores. Figures 2-4 present the scores of the disaggregated parts based on content domain (i.e., life science, physical science, and earth science) respectively. Findings represented by each figure were identified through a series of questions which were posed in order to operationalize the overall question of what are the differences in growth over time between the U.S and selected countries on average correct responses of questions within the same content domains? The questions are as follows: are there nations that consistently outpace all other nations? Are there nations that consistently lag behind all other countries? Are there groups of nations that move between multiple rankings throughout the five iterations of the TIMSS? Is there an overall positive or negative growth pattern of scores over time? Is there a convergence of scores over time? The findings that are presented, based on each figure, have been organized per these five questions. In addition to these questions, I have presented in the summary of this section, findings based on comparisons across content domains as well as across grade level.

Overall science scores. A comparison of growth over time of fourth-grade average scale scores in science for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 1). Singapore's average scale scores are consistently higher than all other nations. There is no nation that lags behind all other nations. Except for Singapore, all other nations have average scale scores which move between multiple rankings throughout the four iterations of the TIMSS that tested fourth-grade students, showing a growth trend that moves up and down the scale. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation, with the exception of Singapore, has a flat growth rate. There is no apparent convergence of scores over time. Singapore's scores diverge positively from the scores of the other nations.

These findings suggest that Singapore is making advances in fourth-grade science scores. Furthermore, all other nations included in the current study are making little or no progress.

Unlike the findings mentioned above, which were formed from scores that contain the overall science scores for the fourth grade, the following three sections of findings were established using the disaggregated scores based on content domain (i.e., life science, physical science, and earth science) respectively. Furthermore, these sections report the findings of the constituent parts of the overall science scores for the fourth grade.

Life science scores. A comparison of growth over time of fourth-grade average scale scores in life science for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 2). Singapore's average scale scores are consistently higher than all other nations. There is no nation that lags behind all other nations. Except for Singapore, all other nations have average scale scores which move between multiple rankings throughout the four iterations of the TIMSS that tested fourth-grade students, showing a growth trend that moves up and down the scale. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation, with the exception of Singapore, has a flat growth rate. There is no apparent convergence of scores over time. Singapore's scores diverge positively from the scores of the other nations.

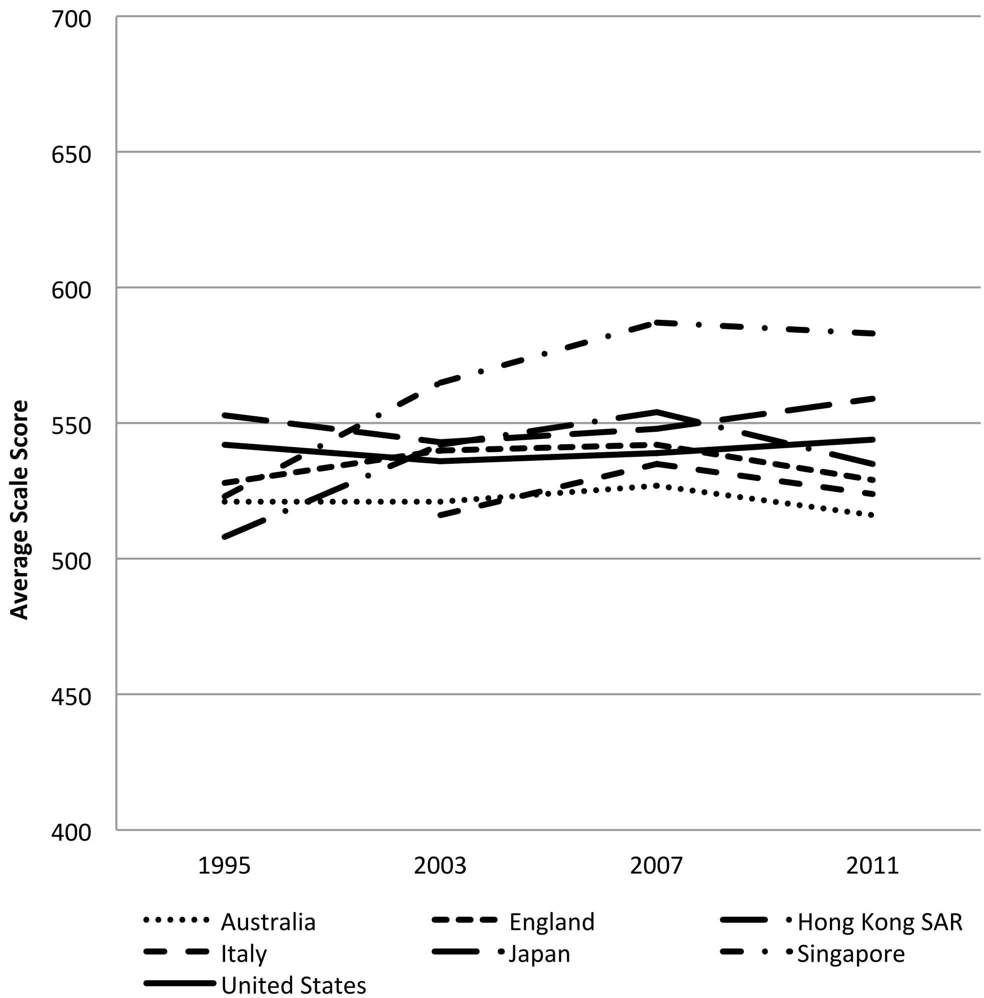


Figure 1. Fourth Grade Average Scale Scores in Science

These findings suggest that Singapore is making growth in fourth-grade life science scores. Furthermore, all other nations included in the current study are making little or no growth.

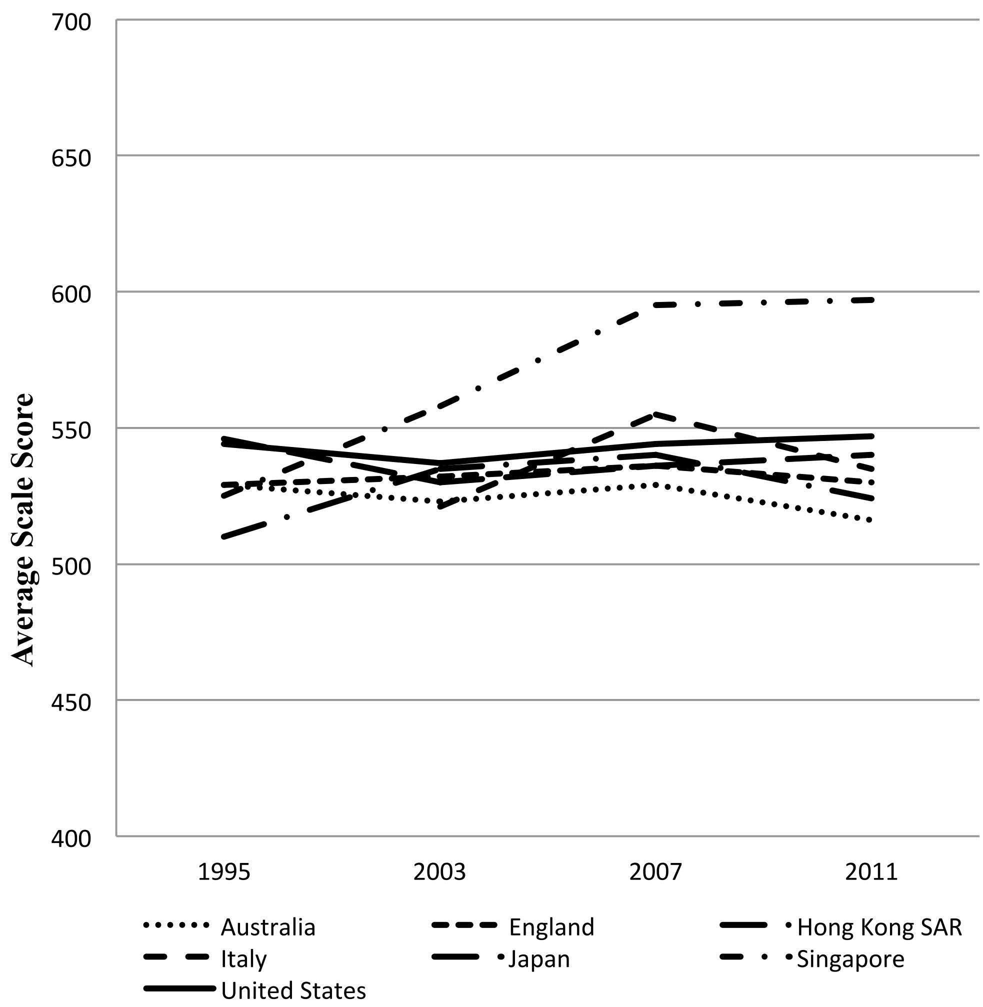


Figure 2. Fourth Grade Average Scale Scores in Life Science

Physical science scores. A comparison of growth over time of fourth-grade average scale scores in the content domain physical science for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 3). Singapore and Japan have average scale scores that are consistently higher than all other nations for all four iterations of the

TIMSS. They trade position in the rankings only after the first year. Italy and Australia have average scale scores that are lower than all other nations score with the exception of the 1995 Hong Kong score, which is lower than Australia's score. The U.S., England, and Hong Kong have average scale scores which move between multiple rankings throughout the four iterations of the TIMSS that tested fourth-grade students, showing a growth trend that moves up and down the scale. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. When looking at the trend lines of Japan and Singapore through the years collectively, there is an apparent positive overall growth. The remaining nations appear to have a flat growth rate. Furthermore, there is no apparent convergence of scores over time.

These findings suggest that Singapore and Japan are making growth in fourth-grade physical science scores. Furthermore, all other nations included in the current study are making little to no growth.

Earth science scores. A comparison of growth over time of fourth-grade average scale scores in the content domain earth science for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 4). There is not a nation that consistently outpaces or lags behind all other nations. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. When looking at the trend lines of Hong Kong and Singapore through the years collectively, there is an apparent positive overall growth, although there is a marked downward trend from 2007 to 2011. Furthermore, there is no apparent convergence of scores over time.

These findings suggest that the U.S growth is similar to the other nations included in the current study. Unlike the findings of the fourth-grade life science and physical science content domain scores, the earth science content domain scores show that the U.S. is neither ahead nor behind, but is instead in line with the other nations included in the current study.

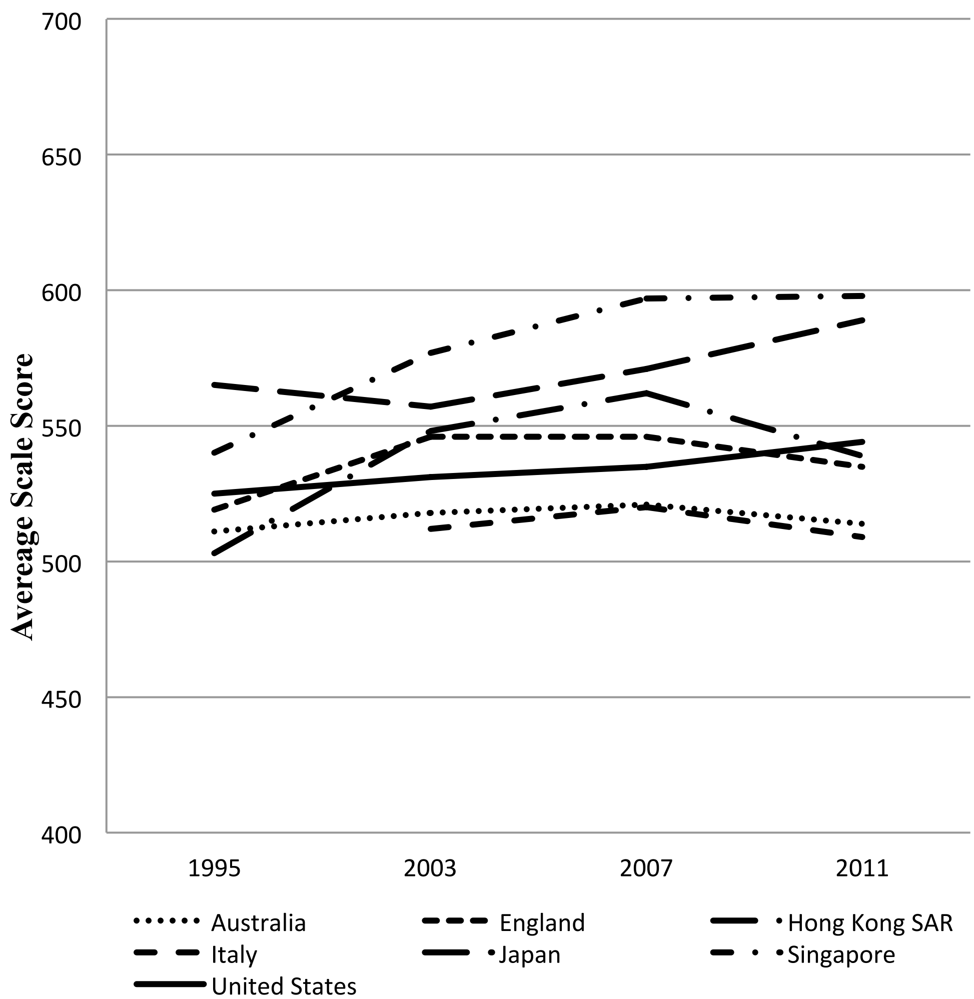


Figure 3. Fourth Grade Average Scale Scores in Physical Science

Eighth-grade science scores. This section contains the figures that present the eighth-grade science scores. Figure 5 presents the overall fourth-grade science scores. Figures 6-9 present the scores of the disaggregated parts based on content domain (i.e., biology, physics, chemistry, and earth science) respectively. Findings represented by each figure were identified through a series of questions which were posed in order to operationalize the overall question of what are the differences in growth over time between the U.S and selected countries on average correct responses of questions within the same content domains? The questions are as follows: are there nations that consistently outpace all other nations? Are there nations that consistently lag behind all other countries? Are there groups of nations that move between multiple rankings throughout the five iterations of the TIMSS? Is there an overall positive or negative growth pattern of scores over time? Is there a convergence of scores over time? The findings that are presented, based on each figure, have been organized per these five questions. In addition to these questions, I have presented in the summary of this section, findings based on comparisons across content domains as well as across grade level.

Overall science scores. A comparison of growth over time of eighth-grade average scale scores in science for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 5). Singapore's average scale scores are consistently higher than all other nations. Japan's average scale scores are also consistently higher than all other nations, with the exception of Singapore for all of the TIMSS iterations, as well as in 2003 when Hong Kong advanced slightly ahead of the other nations included in the current study. Italy has average scale scores that are consistently lower than all other nations. Australia, England, Hong Kong, Japan, and the U.S. have average scale scores which move between multiple rankings throughout the five iterations of the TIMSS that tested eighth-grade students, showing a growth trend that moves up and down the scale. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation has a flat growth rate. Hong Kong's growth spikes in 2003 and then returns to a lower scale score the following years. There is no apparent convergence of scores over time.

These findings suggest that the U.S. is part of a larger group of nations which are consistently in the middle of the pack with regards to average scale score in science. Furthermore, Singapore is consistently leading and Italy is consistently lagging, showing that the U.S. is behind in science overall.

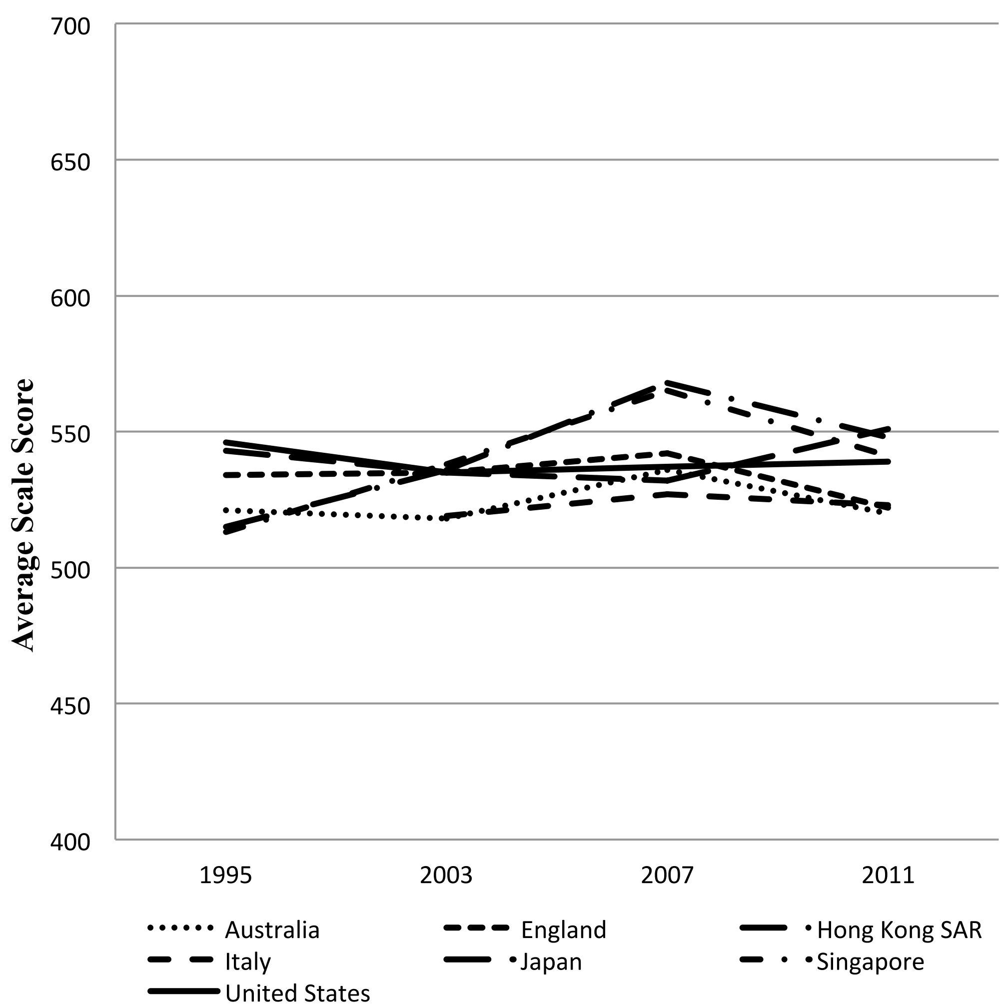


Figure 4. Fourth Grade Average Scale Scores in Earth Science

Unlike the findings mentioned above, which were formed from scores that contain the overall science scores for the eighth grade, the following four sections of findings were established using the disaggregated scores based on content domain (i.e., biology, physics, chemistry, and earth science) respectively. Furthermore, these sections report the findings of the constituent parts of the overall science scores for the eighth grade.

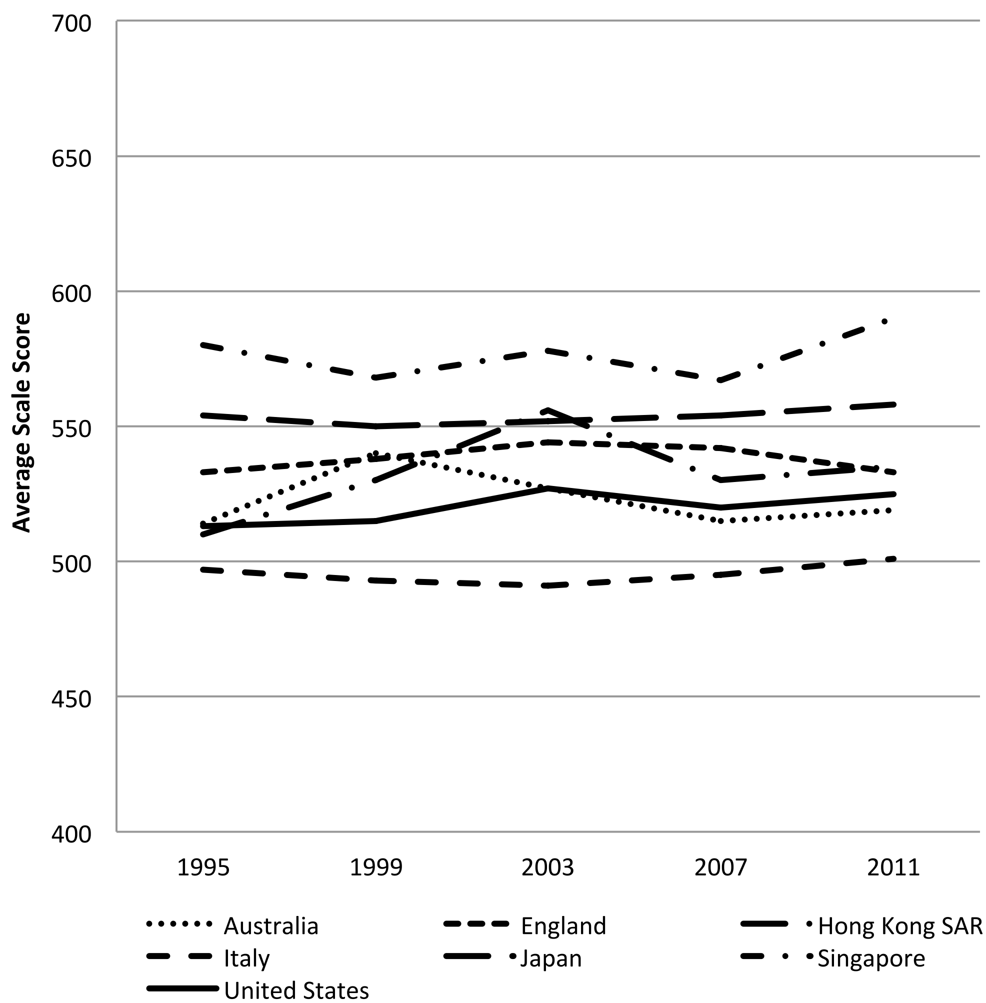


Figure 5. Eighth Grade Average Scale Scores in Science

Biology scores. A comparison of growth over time of eighth-grade average scale scores in the content domain biology for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 6). Singapore's average scale scores are consistently higher than all other nations. Japan's average scale scores are also consistently higher than all other nations, with the exception of Singapore for all of the TIMSS iterations, as well as in 2003 when Hong Kong advanced slightly ahead of the other nations included in the current study. Italy has average scale scores that are consistently lower than all other nations. Australia, England, Hong Kong, Japan, and the U.S. have average scale scores which move between multiple rankings throughout the five iterations of the TIMSS that tested eighth-grade students, showing a growth trend that moves up and down the scale. When looking at the trend lines of nations through the years collectively, there is an apparent positive overall growth. Hong Kong's growth spikes in 2003 and then returns to a lower scale score the following years. There is no apparent convergence of scores over time.

These findings suggest that the U.S. is part of a larger group of nations which are consistently in the middle of the pack with regards to average scale scores in biology. Furthermore, Singapore is consistently leading and Italy is consistently lagging, showing that the U.S. is behind in biology. Also, the apparent positive growth of the scores of all the nations collectively suggests that understanding of biology is increasing across the board.

Physics scores. A comparison of growth over time of eighth-grade average scale scores in the content domain physics for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 7). Unlike the content domains biology, chemistry, and earth science, the trends in average scale scores are more striated. Singapore's average scale scores are, again, consistently higher than all other nations. Japan's average scale scores are consistently lower than Singapore's but higher than all other countries. Italy's average scale scores are consistently lower than all other nations. Australia, England, Hong Kong, and the U.S. have average scale scores which move between multiple rankings throughout the five iterations of the TIMSS that tested eighth-grade students, showing a growth trend that moves up and down the scale. Although, unlike other content domain scores, the U.S. changes rank only once in 2011 and trades position in the rankings with Australia. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation has a flat growth rate. There is no apparent convergence of scores over time.

These findings suggest that for eighth-grade physics average scale scores for each nation show a marked difference between each other with the exception of Hong Kong and England which exchange rank between each other multiple times. Furthermore, it can be seen that the U.S. is lagging behind in eighth-grade physics and is not making any growth.

Chemistry scores. A comparison of growth over time of eighth-grade average scale scores in the content domain chemistry for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 8). Singapore's average scale scores are consistently higher than all other nations. Japan's average scale scores are also consistently higher than all other nations, with the exception of Singapore for all of the TIMSS iterations. Italy has average scale scores that are consistently lower than all other nations. Australia, England, Hong Kong, and the U.S. have average scale scores which move between multiple rankings throughout the five iterations of the TIMSS that tested eighth-grade students, showing a growth trend that moves up and down the scale. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Some nations (i.e., Australia and Italy) have a negative growth rate, others (i.e., Japan and Singapore) have a positive growth rate, and the rest (i.e., Australia, England, Hong Kong, and the U.S.) have a flat growth rate. Hong Kong's growth spikes in 2003 and then returns to a lower scale score the following years. There is no apparent convergence of scores over time, instead it is apparent that an overall divergent trend can be seen.

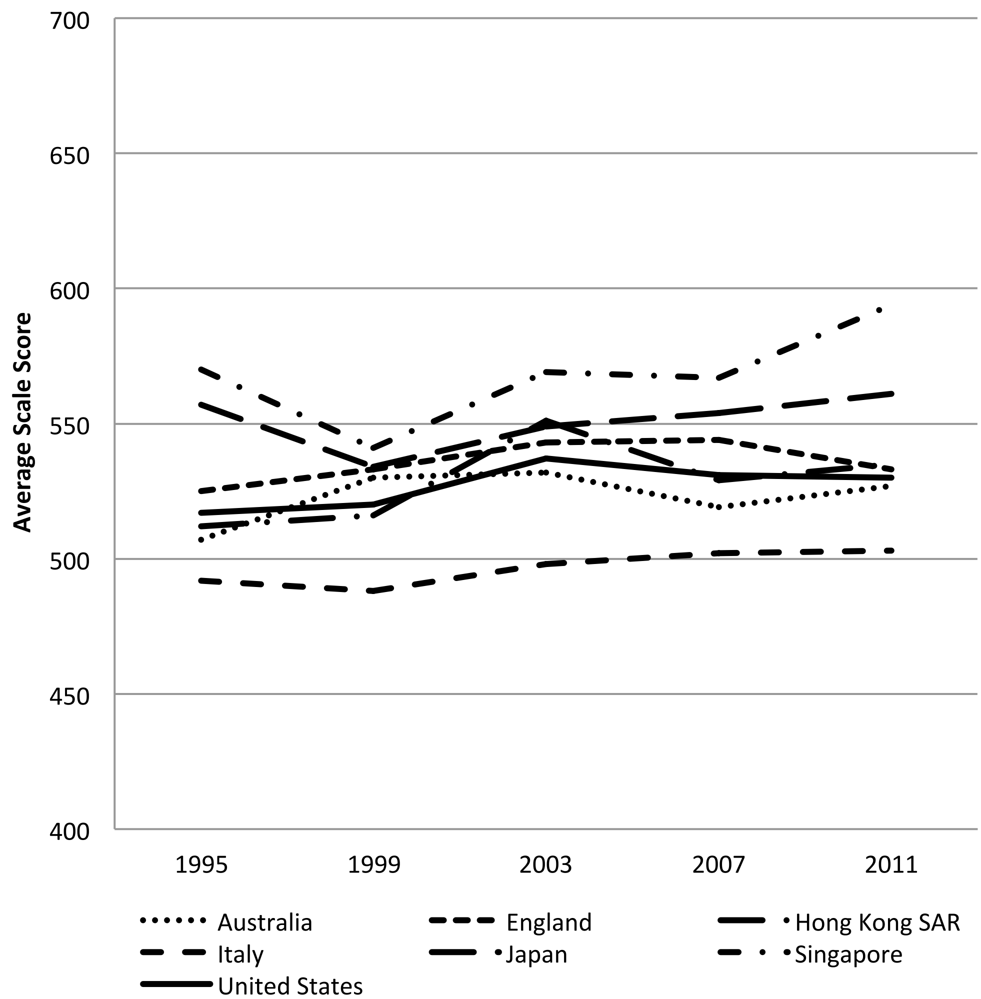


Figure 6. Eighth Grade Average Scale Scores in the Content Domain Biology

These findings suggest that the U.S. is part of a larger group of nations which are consistently in the middle of the pack with regards to average scale score in science. Furthermore, Singapore is consistently leading and Italy is consistently lagging; the U.S. is behind in chemistry and is closer to the Italy scores than to the Singapore scores. Also, it is apparent that some nations are making growth while others are remaining steady and still others are digressing in chemistry scores.

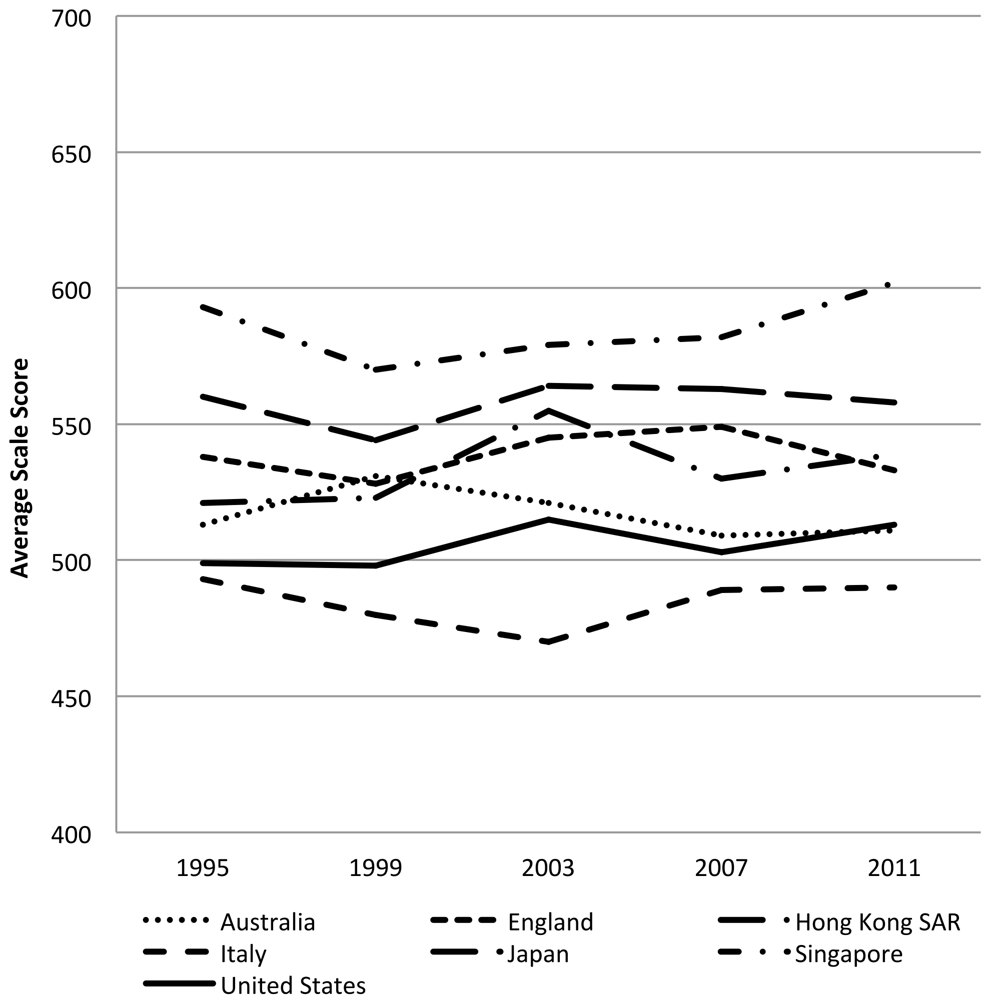


Figure 7. Eighth Grade Average Scale Scores in Physics

Earth science scores. A comparison of growth over time of eighth-grade average scale scores in the content domain earth science for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 9). Earth science scores show the most unique findings in comparison to the content domains biology, physics and chemistry. All nations have average scale scores which move between multiple rankings throughout the five iterations of the TIMSS that tested eighth-grade students, showing a growth trend that moves up and down the scale. When looking at the trend lines of nations through the years collectively, there is an apparent positive overall growth. Convergence of scores over time is seen between the U.S., Hong Kong, Japan, England, and Australia. The average scale scores for Singapore and Italy are diverging positive and negative, respectively over time.

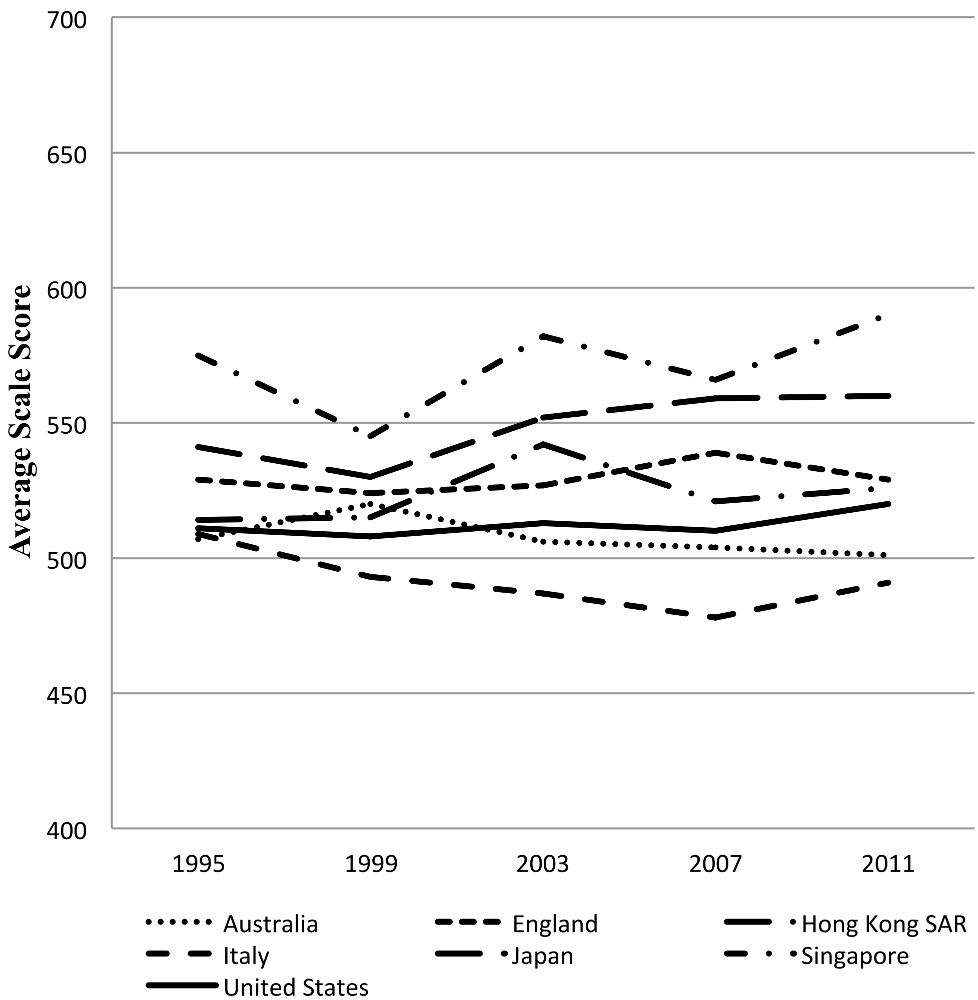


Figure 8. Eighth Grade Average Scale Scores in Chemistry

The findings suggest that the growth in average scale scores of the U.S., Hong Kong, Japan, England, Singapore, and Australia are similar in their upward trend as well as their convergence. The U.S. appears to be in line with other nations with regards to earth science.

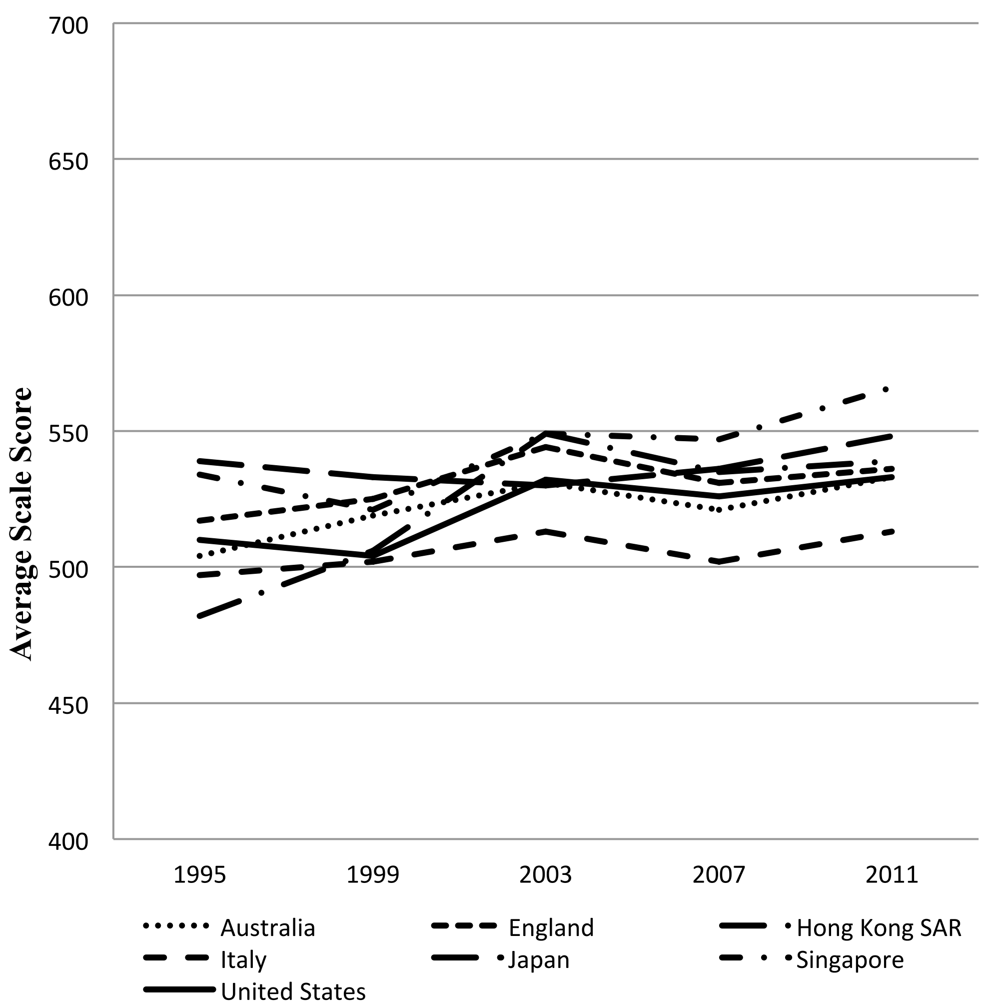


Figure 9. Eighth Grade Average Scale Scores in Earth Science

Fourth-grade mathematics scores. This section contains the figures that present the fourth-grade mathematics scores. Figure 10 presents the overall fourth-grade science scores. Figures 11-13 present the scores of the disaggregated parts based on content domain (i.e., number, geometry, and data) respectively. Findings represented by each figure were identified through a series of questions, which were posed in order to operationalize the overall question of what are the differences in growth over time between the U.S and selected countries on average correct responses of questions within the same content domains? The questions are as follows: are there nations that consistently outpace all other nations? Are there nations that consistently lag behind all other countries? Are there groups of nations that move between multiple rankings throughout the five iterations of the TIMSS? Is there an overall positive or negative growth pattern of scores over time? Is there a convergence of scores over time? The findings that are presented, based on each figure, have been organized per these five questions. In addition to these questions, I have presented in the summary of this section, findings based on comparisons across content domains as well as across grade level.

Overall mathematics scores. A comparison of growth over time of fourth-grade average scale scores in mathematics for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 10). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia. When looking at the trend lines of nations through the years collectively, there is an apparent positive overall growth. There is no apparent convergence of scores over time.

These findings suggest that Hong Kong, Japan, and Singapore excel in mathematics in comparison to the U.S., England, Italy, and Australia for the fourth grade.

Unlike the findings mentioned above, which were formed from scores that contain the overall mathematics scores for the fourth grade, the following three sections of findings were established using the disaggregated scores based on content domain (i.e., number, geometry, and data) respectively. Furthermore, these sections report the findings of the constituent parts of the overall mathematics scores for the fourth grade.

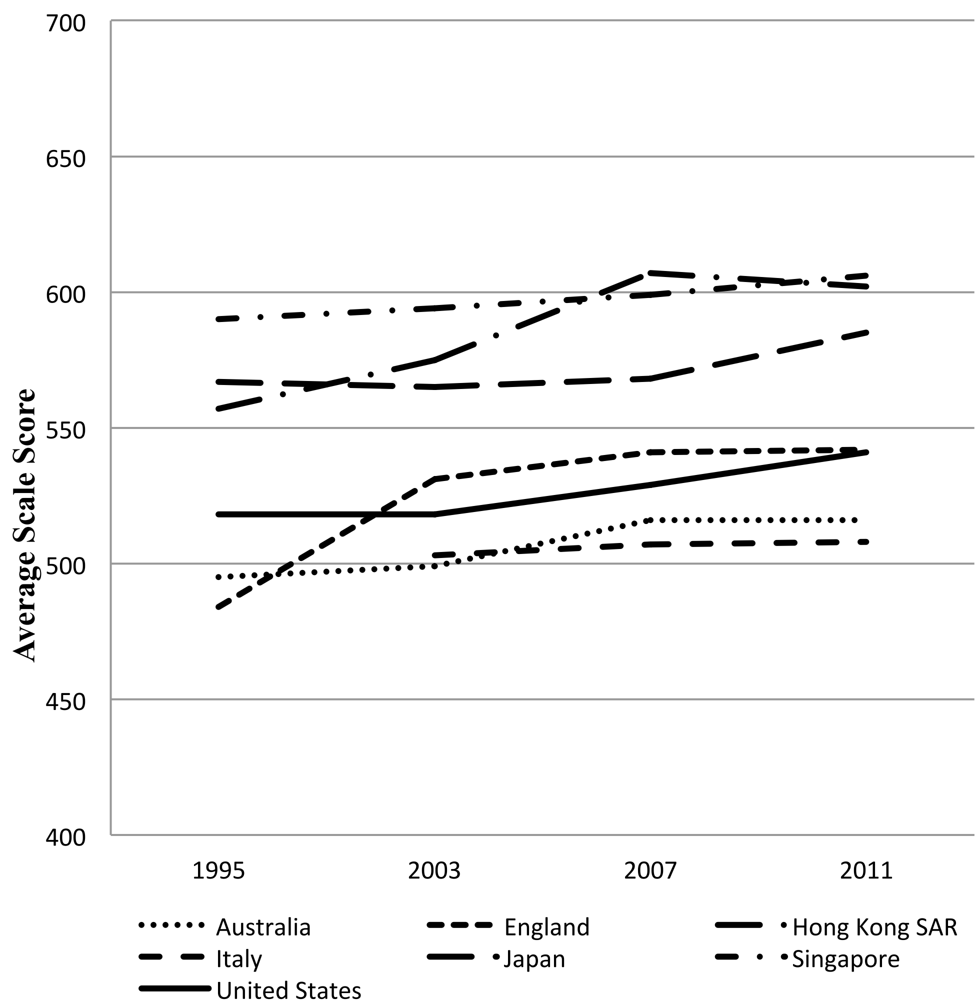


Figure 10. Fourth Grade Average Scale Scores in Mathematics

Number scores. A comparison of growth over time of fourth-grade average scale scores in mathematics content domain number for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 11). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia.

When looking at the trend lines of nations through the years collectively, there is an apparent positive overall growth. There is no apparent convergence of scores over time.

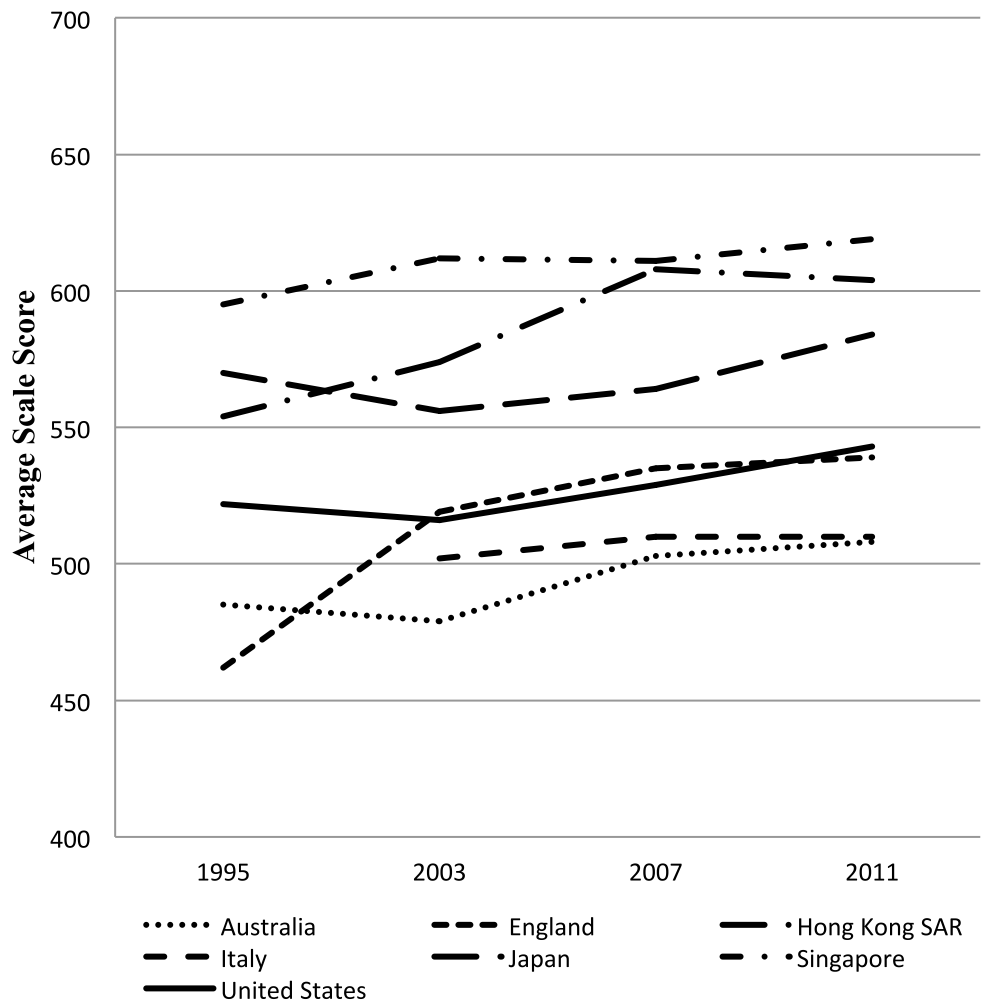


Figure 11. Fourth Grade Average Scale Scores in Number

The findings represented in Figure 11 suggest that Hong Kong, Japan, and Singapore excel in mathematics content domain Number in comparison to the U.S., England, Italy, and Australia for the fourth grade.

Geometry scores. A comparison of growth over time of fourth-grade average scale scores in mathematics content domain geometry for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 12). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia. When looking at the trend lines of nations through the years collectively, there is an apparent positive overall growth. There is no apparent convergence of scores over time.

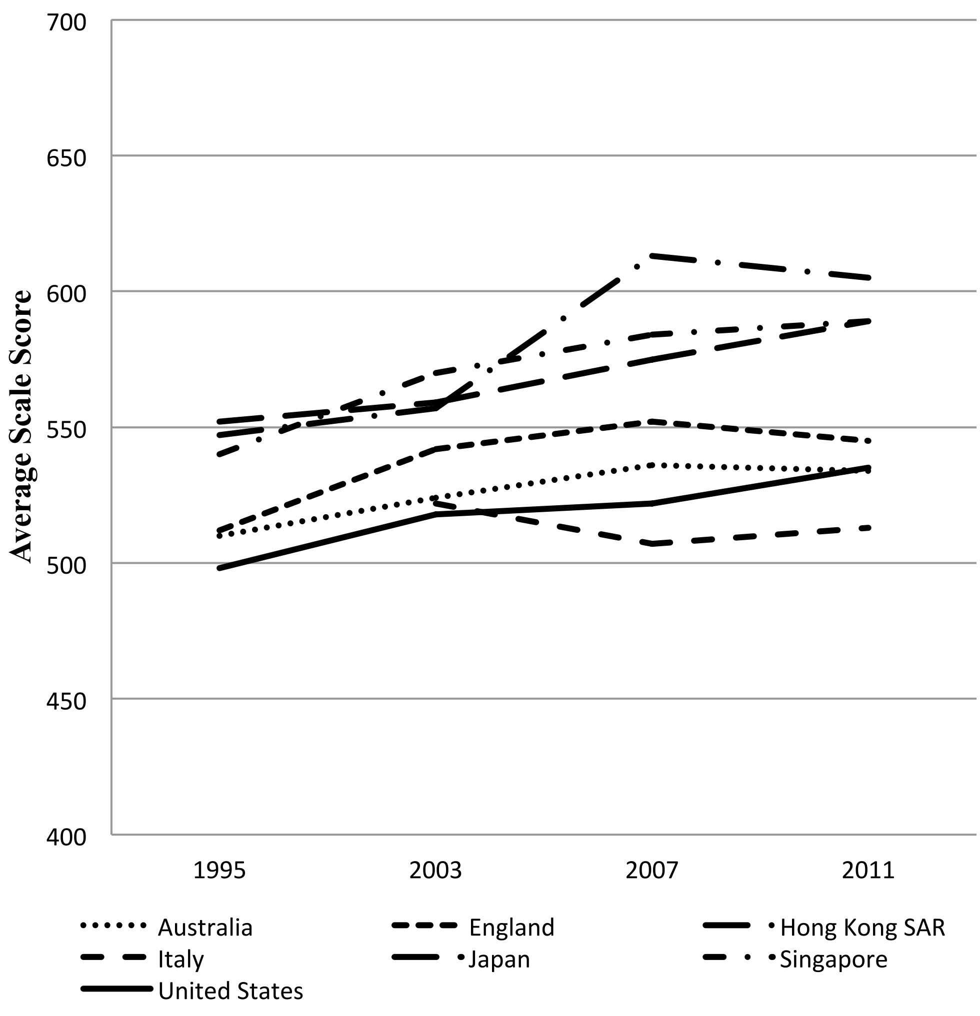


Figure 12. Fourth Grade Average Scale Scores Geometry

The findings represented in Figure 12 suggest that Hong Kong, Japan, and Singapore excel in mathematics content domain geometry in comparison to the U.S., England, Italy, and Australia for the fourth grade.

Data scores. A comparison of growth over time of fourth-grade average scale scores in mathematics content domain data for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 13). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia as time goes on. The difference in scores between eastern and western nations is not as apparent in the first two iterations of the TIMSS, but by 2007 the division between eastern and western nations becomes defined. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. When looking at the trend lines of eastern nations through the years collectively, there is an apparent positive overall growth. Western nations, on the other hand show only a flat growth pattern. There is no apparent convergence of scores over time in western nations, but eastern nations show convergence over time. By 2011 the scores of eastern nations are very similar.

The findings represented in Figure 13 suggest that Hong Kong, Japan, and Singapore excel in mathematics content domain data in comparison to the U.S., England, Italy, and Australia for the fourth grade. Furthermore, eastern nations have grown overall in data while western nations have experienced a flat growth pattern.

Eighth-grade mathematics scores. This section contains the figures that present the eighth-grade mathematics scores. Figure 14 presents the overall eighth-grade mathematics scores. Figures 15-18 present the scores of the disaggregated parts based on content domain (i.e, number, geometry, data, and algebra) respectively. Findings represented by each figure were identified through a series of questions, which were posed in order to operationalize the overall question of what are the differences in growth over time between the U.S and selected countries on average correct responses of questions within the same content domains? The questions are as follows: are there nations that consistently outpace all other nations? Are there nations that consistently lag behind all other countries? Are there groups of nations that move between multiple rankings throughout the five iterations of the TIMSS? Is there an overall positive or negative growth pattern of scores over time? Is there a convergence of scores over time? The findings that are presented, based on each figure, have been organized per these five questions. In addition to these questions, I have presented in the summary of this section, findings based on comparisons across content domains as well as across grade level.

Overall mathematics scores. A comparison of growth over time of eighth-grade average scale scores in mathematics for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 14). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation has a flat growth rate. There is no apparent convergence of scores over time.

These findings suggest that Hong Kong, Japan, and Singapore excel in mathematics in comparison to the U.S., England, Italy, and Australia. The division of eastern and western nations varies in severity depending on the content domain.

Unlike the findings mentioned above, which were formed from scores that contain the overall mathematics scores for the eighth grade, the following four sections of findings were established using the disaggregated scores based on content domain (i.e., number, geometry, data, and algebra) respectively. Furthermore, these sections report the findings of the constituent parts of the overall mathematics scores for the eighth grade.

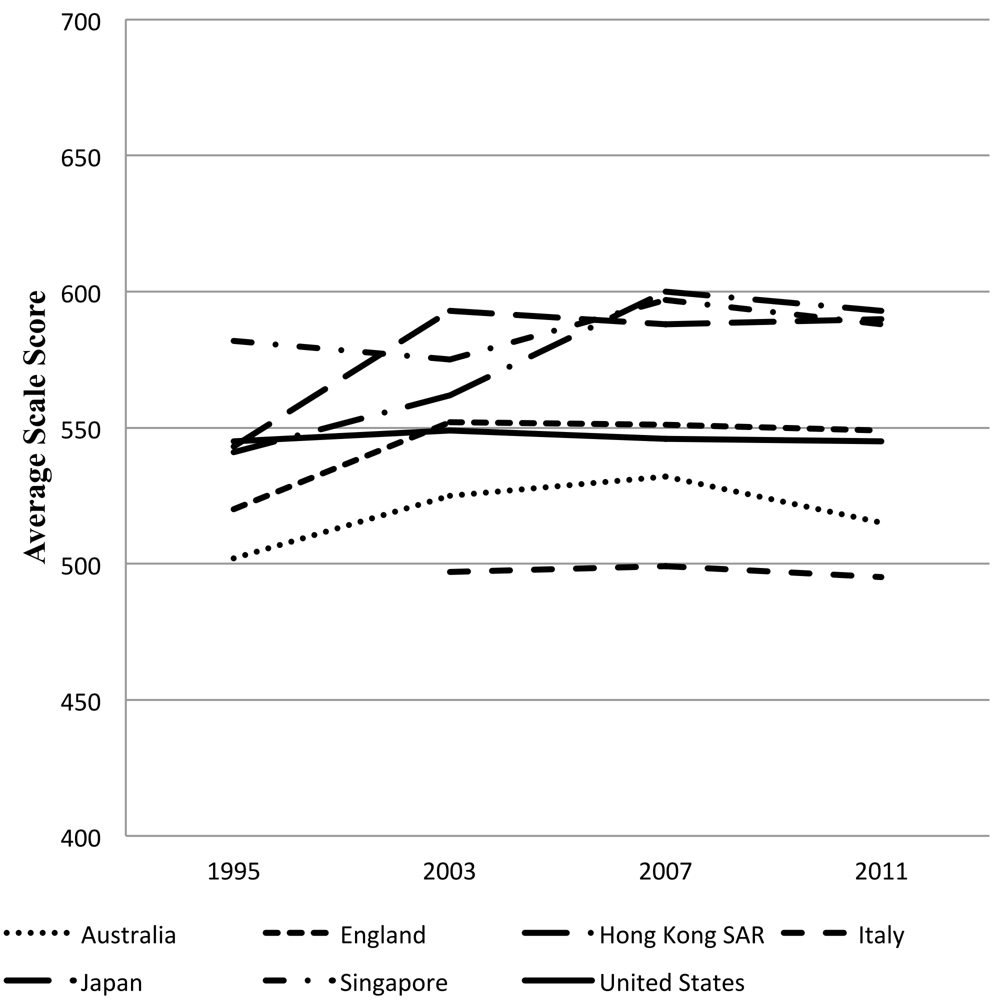


Figure 13. Fourth Grade Average Scale Scores in Data

Number scores. A comparison of growth over time of eighth-grade average scale scores in the content domain number for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 15). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia.

When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation has a flat growth rate. There is no apparent convergence of scores over time.

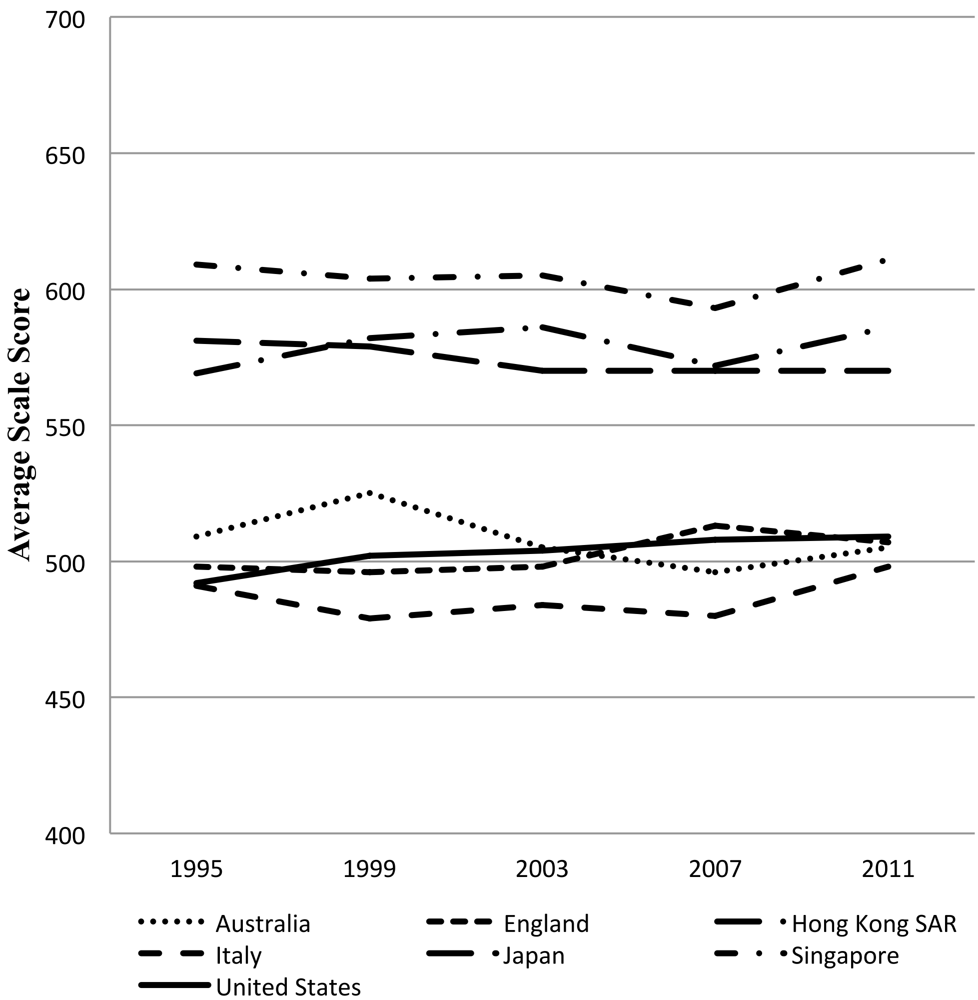


Figure 14. Eighth Grade Average Scale Scores in Mathematics

The findings represented in Figure 15 suggest that Hong Kong, Japan, and Singapore excel in the content domain number in comparison to the U.S., England, Italy, and Australia. The division between eastern and western nations for the content domain number is not as apparent as with geometry and algebra.

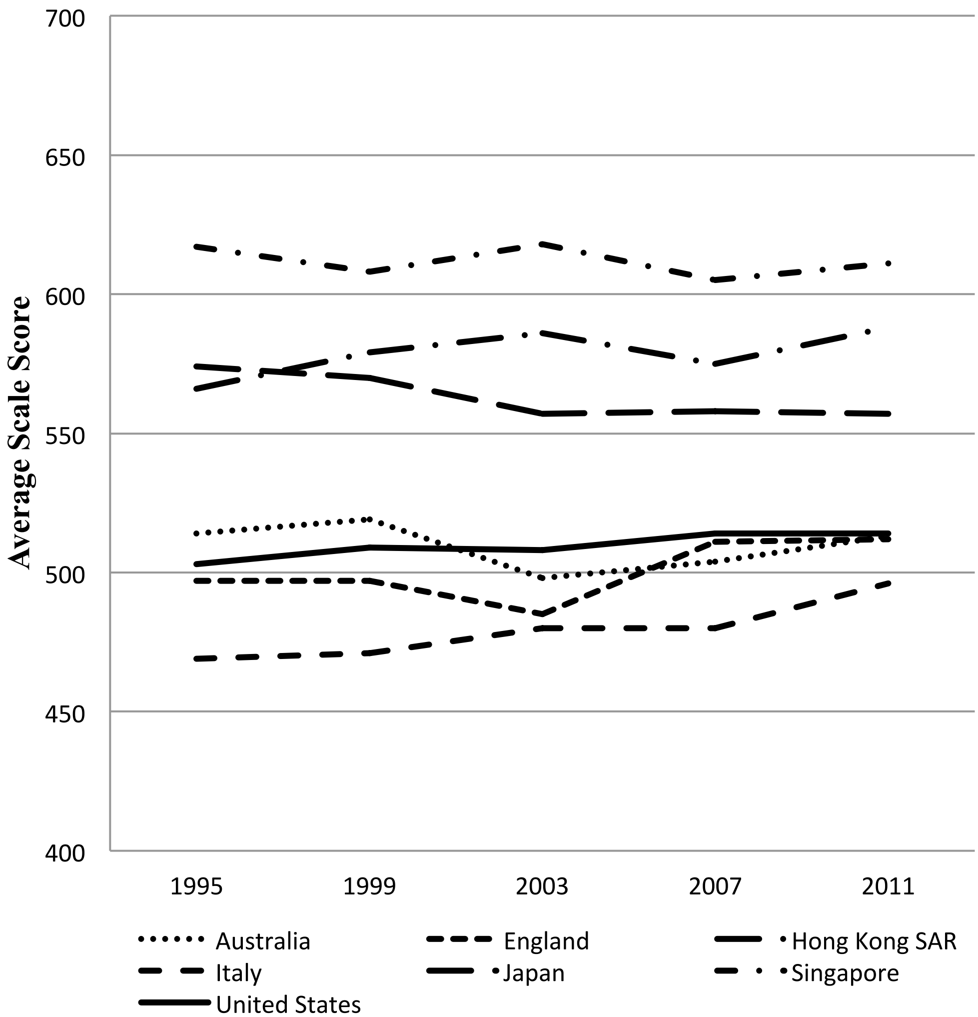


Figure 15. Eighth Grade Average Scale Scores in Number

Geometry scores. A comparison of growth over time of eighth grade average scale scores in the content domain geometry for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 16). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia.

When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation has a flat growth rate. There is no apparent convergence of scores over time.

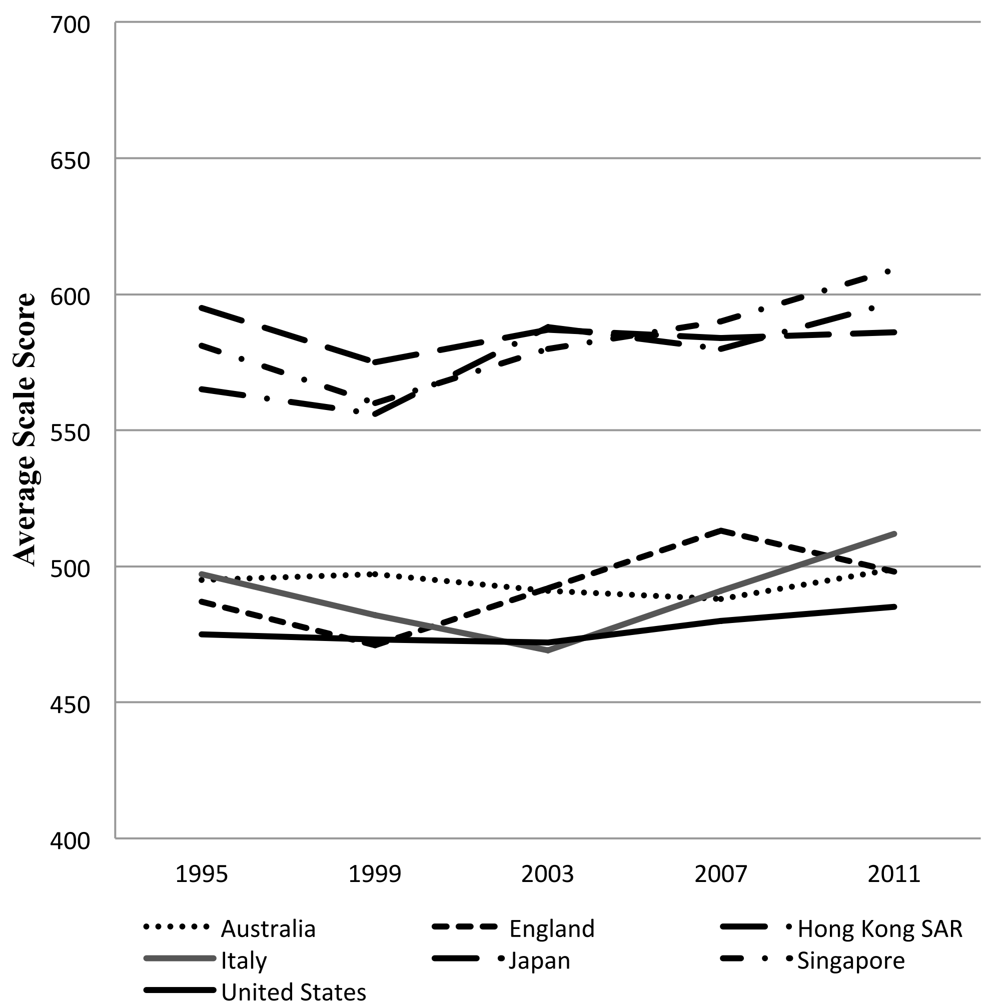


Figure 16. Eighth Grade Average Scale Scores in Geometry

The findings represented in Figure 16 suggest that Hong Kong, Japan, and Singapore excel in the content domain geometry in comparison to the U.S., England, Italy, and Australia.

The division between eastern and western nations is very defined, unlike the content domains number and data.

Data scores. A comparison of growth over time of eighth-grade average scale scores in the content domain data for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 17). Hong Kong, Japan, and Singapore, and Italy have average scale scores that have a consistent rank among the countries scores. Australia, England and the U.S. have average scale scores that move between fourth and sixth. When looking at the trend lines of nations through the years collectively, there is an apparent positive overall growth although Italy has a trend line that is flat in comparison to the other nations. There is no apparent convergence of scores over time.

The findings represented in Figure 17 suggest that Hong Kong, Japan, and Singapore excel in the content domain data in comparison to the U.S., England, Italy, and Australia and that Italy is lagging behind. Similar to the content domain number, the division between eastern and western nations for the content domain data is not as apparent as with geometry and algebra.

Algebra scores. A comparison of growth over time of eighth-grade average scale scores in the content domain algebra for the U.S., Hong Kong, Japan, Singapore, England, Italy, and Australia, shows multiple findings (see Figure 18). Hong Kong, Japan, and Singapore have average scale scores that are consistently higher than the U.S., England, Italy, and Australia. When looking at the trend lines of nations through the years collectively, there is no apparent positive or negative overall growth. Each nation has a flat growth rate. There is no apparent convergence of scores over time.

The findings represented in Figure 18 suggest that Hong Kong, Japan, and Singapore excel in the content domain algebra in comparison to the U.S., England, Italy, and Australia.

Similar to the content domain geometry, the division between eastern and western nations is very defined, unlike the content domains number and data.

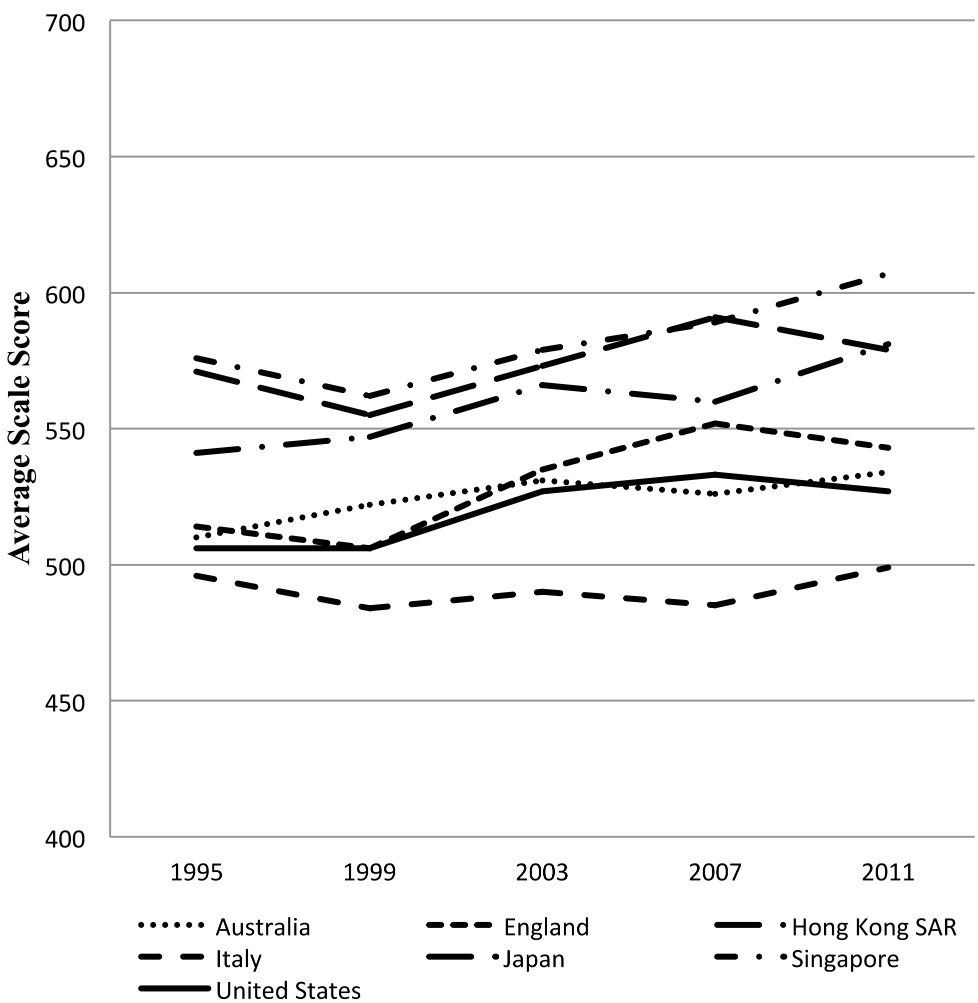


Figure 17. Eighth Grade Average Scale Scores in Data

Summary of findings for research question 1. Analysis for question 1 revealed very little growth over the history of the TIMSS for the nations included in the current study. Most nations were consistent with their scores over time. Even though many of the nations had a large range of scores, they did not grow consistently. For example sometimes a nation's scores would begin with low scores and then receive higher scores the following year and would finally end slightly higher or slightly lower than the original score. This pattern made the overall growth of the nations' scores negligible. Gains and losses could only been seen between individual years and not over all the iterations of the TIMSS. There are some exceptions to this pattern. For the fourth-grade mathematics content domains of geometry and number and eighth-grade mathematics content domain data, all nations average scale scores had a positive trend over time. Singapore's average scale scores had a positive trend over time for fourth-grade science content domains of physical science and life science as well as for the fourth-grade overall science scores over time. Japan's average scale scores had a positive trend over time for fourth-grade science content domain physical science.

When examining average scale scores across a range of content domains, it is clear that the scores seem to be very similar between all nations included in the current study. Findings for eighth-grade overall science scores do not show any large gaps between the scores of the different nations. This is most evident when looking at Figure 9. This same pattern can be seen when looking at findings for fourth-grade science (see Figures 1-4) with the exception of the Singapore scores that appear to make gains by the 2007 and 2011 iterations of the TIMSS.

The mathematics scores tell a different story compared to the science scores. Two distinct groups of countries can be seen for all of the mathematics content areas for both fourth and eighth grades. Although the difference in the two groups of countries seems to be more prominent for the eighth grade than the fourth grade. Singapore, Japan and Hong Kong all show scores that appear to be in a group of their own, which are overall higher scores than the U.S., Australia, Italy, and England. Also, unlike the science scores, nearly all of the mathematics scores seem to be growing across time.

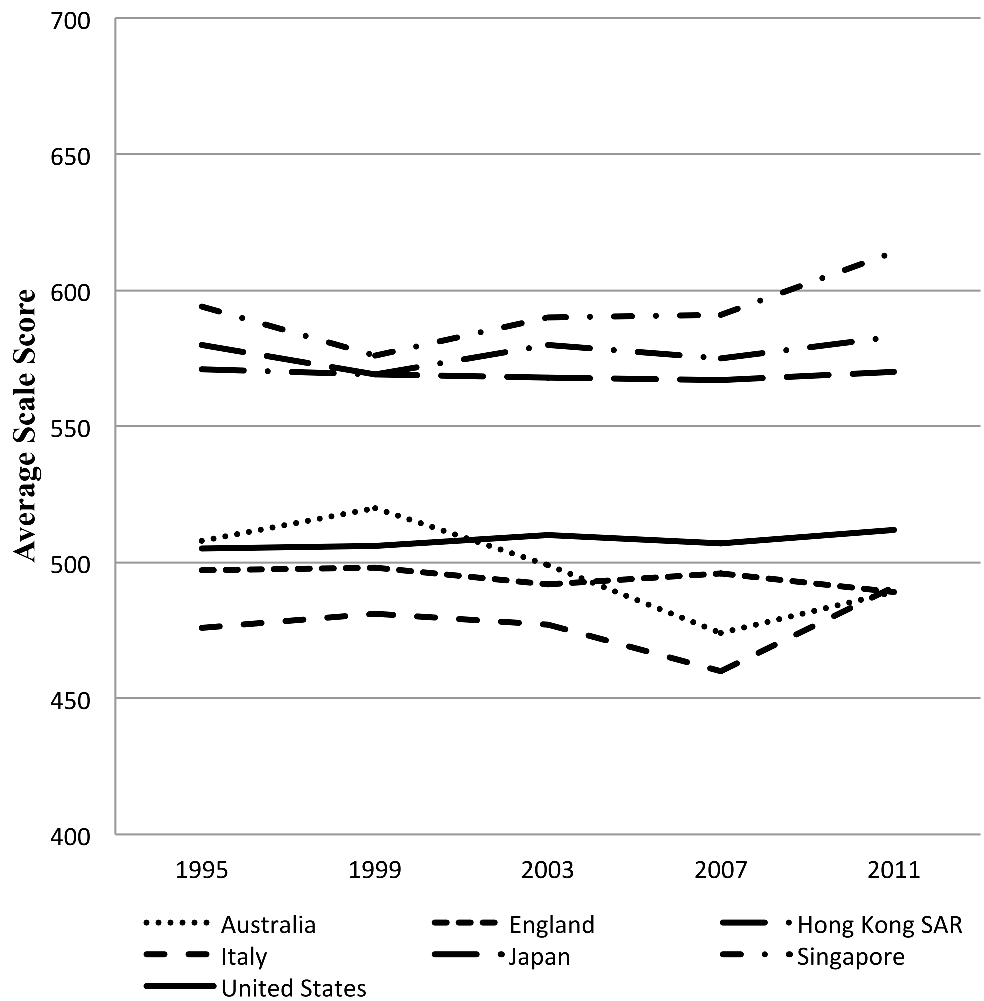


Figure 18. Eighth Grade Average Scale Scores in Algebra

Research Question 2: Comparison Between the U.S. and Selected Countries Across Cognitive Domains

Science cognitive domain comparisons. Multiple independent-samples *t*-tests were conducted to compare the percent of correct student response on individual TIMSS science questions based on cognitive domain between the U.S. and six participating nations, namely

Australia, England, Hong Kong, Italy, Japan, and Singapore (see Table 3). There was a significant difference in the scores of the U.S. and the scores of Singapore for the mathematics cognitive domain of reasoning for the fourth grade as well as a significant difference of the science cognitive domains of knowing, applying, and reasoning for the eighth grade with Singapore having higher scores in all areas of significant difference. There was a significant difference in the scores of the U.S. and the scores of Japan for the science cognitive domain of applying for the eighth grade with Japan having higher scores.

Mathematics cognitive domain comparisons. Multiple independent-samples *t*-tests were conducted to compare the percent of correct student response on individual TIMSS mathematics questions based on cognitive domain between the U.S. and six participating nations, namely Australia, England, Hong Kong, Italy, Japan, and Singapore (See, Table 4) There was a significant difference between the scores of the U.S. and the scores of Singapore for the mathematics cognitive domains of knowing, applying and reasoning for both fourth and eighth grades with Singapore having higher scores in all areas of significant difference. There was a significant difference between the scores of the U.S. and the scores of Hong Kong for the mathematics cognitive domains of knowing and applying for the fourth grade and the mathematics cognitive domains of knowing, applying, and reasoning for the eighth grade with Hong Kong having higher scores in all areas of significant difference. There was a significant difference in the scores of the U.S. and the scores of Japan for the mathematics cognitive domains of applying and reasoning for the eighth grade with Japan having higher scores. Italy and Australia both showed a significant difference between the U.S. for the mathematics cognitive domain of knowing with the U.S. having higher scores in both instances.

Table 3

P-values Comparing the Difference Between US and Selected Country Scores for Science Across Cognitive Domains.

	US and Australia	US and England	US and Hong Kong	US and Italy	US and Japan	US and Singapore
			Fourth Grade			
Knowing	0.48	0.69	0.66	0.97	0.71	0.26
Applying	0.28	0.64	0.55	0.96	0.66	0.22
Reasoning	0.18	0.74	0.54	0.26	0.16	0.02*
			Eighth Grade			
Knowing	0.58	0.63	0.27	0.44	0.47	0.01*
Applying	0.69	0.73	0.62	0.32	0.03*	0.01*
Reasoning	0.63	0.68	0.86	0.05	0.13	0.01*

Note. * = $p < .05$ showing a significant difference in which the country being compared to the US has scores that are significantly higher.

Summary of findings for research question 2. Analysis for question 2 revealed that the U.S. percent student correct response of individual questions were significantly different in a number of cognitive domains for science and mathematics for several countries. Singapore students are more skilled for all but two of the cognitive domains (i.e., knowing and applying in science for the fourth grade). Japan and Hong Kong students are more skilled on a number of cognitive domains for both science and mathematics for the fourth and eighth grades (i.e., Japan: applying in science for the eighth grade and applying and reasoning in mathematics for the eighth grade; Hong Kong: knowing and applying in mathematics for the fourth grade and knowing, applying, and reasoning in mathematics for the eighth grade).

Table 4

P-values Comparing the Difference Between US and Selected Country Scores for Mathematics Across Cognitive Domains.

	US and Australia	US and England	US and Hong Kong	US and Italy	US and Japan	US and Singapore
			Fourth Grade			
Knowing	0.01**	0.37	0.00*	0.00**	0.99	0.00*
Applying	0.43	0.88	0.02*	0.13	0.11	0.00*
Reasoning	0.47	0.89	0.05	0.44	0.11	0.02*
			Eighth Grade			
Knowing	0.29	0.20	0.00*	0.17	0.05	0.00*
Applying	0.85	0.74	0.00*	0.71	0.00*	0.00*
Reasoning	0.61	0.35	0.00*	0.91	0.00*	0.00*

Note. * = $p < .05$ showing a significant difference in which the country being compared to the US has scores that are significantly higher, ** = $p < .05$ showing a significant difference in which the country being compared to the US has scores that are significantly lower.

These findings suggest that the U.S. is lagging behind in all science cognitive domains for the eighth grade as well as in the science cognitive domain reasoning for the fourth grade compared to the top performing nation (i.e., Singapore). They also suggest that the U.S. is lagging behind in all mathematics cognitive domains for both the fourth and eighth grades compared to the top performing nation (i.e., Singapore).

Japan did not show any significant difference in the fourth grade but when comparing eighth-grade students Japan scored significantly higher in the applying section of science and the applying and reasoning sections of mathematics. Hong Kong showed a significant difference in knowing and applying in mathematics for the fourth grade and then when comparing eighth-grade students; and Hong Kong showed a significant difference in knowing, applying, and reasoning sections of mathematics. These findings suggest that the U.S. is making less progress from the fourth to eighth grades in both science and mathematics, as can be seen by the larger number of nations which show a significant difference in eighth grade scores than in fourth grade scores.

Chapter 5

Discussion

International educational studies, including the TIMSS, can hold a vast amount of data to be mined. Through 16 years of data collection, TIMSS researchers have created an enormous treasure trove of data on a large portion of the world's nations. With those data, many researchers have attempted to make conclusions about the effectiveness of certain education systems. News organizations from around the country have often cited the high rank of certain countries on international assessments as a reason why the U.S. needs to think about changing its education system (see, Arenson, 2004; Armario, 2010; Asimov, 2000; Ward, & Stewart, 2012;). They cite countries such as Finland and Singapore as examples of education systems that should be mimicked and modeled after (i.e., Robelen, & Gewertz, 2013; Cardno, 2012). Oftentimes the articles only mention the relative rank of these countries as the main reason why we should follow in their footsteps. Many policy makers will also use rank alone as a determining factor for policy (Nagesh, 2010). The results do suggest that the overall ranking of nations is far too simplistic an interpretation of the data collected by the TIMSS.

The picture that is painted by the relative rank of the U.S. in comparison to other nations is often a bleak. In the current study my purpose was to discover whether relative rank is an effective way to compare nations' complex education systems or if there are more useful methods of comparing nations' education systems.

Conclusions and Implications

Graphing the average scale score based on content domain over the course of the TIMSS allowed me to present many findings that helped to answer the question: What are the differences in growth over time? I was also able to present a number of findings based on the series of *t*-tests performed. By combining the findings from each of these two methodologies, I was able to make three overarching conclusions that seemed to encompass all that I found. First, the U.S. is lagging behind eastern nations in mathematics content and cognitive domains. Second, the U.S. is lagging behind Singapore (a top performing nation) in some, but not all science content and cognitive domains. Third, achievement gaps between eastern and western nations increases from the fourth to eighth grade. In addition to these three conclusions, I also found that if standard error were taken into account when reporting rank, then the U.S. would be tied with many nations, including some of the top performing nations, in some of the content and cognitive domains within science and mathematics.

The U.S. is lagging behind Singapore (a top performing nation) in some, but not all science content and cognitive domains. There appeared to be segregation between Singapore and all other nations for many of the science content domains for the fourth and eighth grades. An exception is found in the content domain physical science for the fourth grade in which Singapore and Japan have made marked advances over the years. There were also some content domains that did not show any one nation that was leading or lagging behind. This can be seen for the content domain of earth science for the fourth and eighth grades. The U.S. may not be far behind many nations in science but they are most definitely behind Singapore in many of the science content domain. The U.S. is not behind Singapore or any other nation in the content domain of earth science, but is instead running with the pack.

The U.S. is lagging behind eastern nations in mathematics content and cognitive domains. Findings based on the fourth- and eighth-grade mathematics content domain figures show the greatest disparity between the U.S. and eastern nations who were included in the current study. Nearly every figure showed a distinct segregation between the eastern and western nations. The only exception included the fourth-grade content domain data.

Singapore was significantly different from the U.S. for the science cognitive domain of reasoning for the fourth grade in all of the cognitive domains for the eighth grade, and was significantly different for all of the mathematics cognitive domains for the fourth grade. This suggests that Singapore outperforms the U.S. in nearly every cognitive domain in science and mathematics for the fourth and eighth grade. The only areas in which the Singapore is not significantly different from the U.S. is in fourth-grade science cognitive domains knowing and applying. The U.S. was also outperformed by Hong Kong in fourth-grade math cognitive domains of knowing and applying, as well as eighth-grade math cognitive domains of knowing, applying, and reasoning. Japan outperformed the U.S. in eighth-grade science cognitive domain applying and math cognitive domains applying and reasoning. The conclusion made from these findings is that the U.S., as well as the other western nations, is lagging behind their eastern counterparts in most mathematics content and cognitive domains.

Achievement gaps between eastern and western nations increases from the fourth to eighth grade. As previously mentioned, there is an achievement gap between eastern and western nations in both science and mathematics. The achievement gap is more pronounced in mathematics. It can also be seen that segregation of eastern and western nations’ scores grows larger and more distinct from the fourth to eighth grade. This suggests that while eastern students are advancing in mathematics from the fourth to eighth grade, western students are not making the same gains. The fact that eastern nations outperform the U.S. was also suggested by Robelen, & Gewertz (2013) and Shen (2005).

If standard error where taken into account when reporting rank, then the U.S. would be tied with many nations, including some of the top performing nations, in some of the content and cognitive domains within science and mathematics. This can be illustrated by thinking of international tests, such as the TIMSS, as a track and field meet. During a track and field meet a number of schools in a similar geographical region compete in a number of sporting events. Each event has participants from each school, and each event produces a rank for each school. When the meet is concluded, the overall rank of each school is tabulated and a winning school is crowned. The highest ranking school may not have won every event in which they participated, but they did win many of them. Some of the lowest ranked schools may also have won 1st place in some of the events in the meet, but they still received a low rank overall. Using these rankings, each school can focus on the sporting event that they performed poorly in when they continue their training regimen. With the above illustration, Singapore is the nation that is the overall winner. Even though they did not win 1st place in every event, they did win most of them. The U.S. did not win all of the events, but they did hold their own in fourth-grade science cognitive domains. Only Hong Kong and Singapore took the lead from the U.S. in fourth-grade math cognitive domains. Please note that in the above illustration, a finding of no significant difference would be considered a tie. The U.S. tied more often than not with Japan and Hong Kong, but only tied with Singapore in two of the cognitive domains.

How do these findings affect the purpose of this study, which was to investigate the effectiveness of relative rank of nations on an international test? The findings in the current study help to uncover what the relative rank of the U.S. means. Even though the U.S. may be ranked seventh in fourth-grade science for 2011 TIMSS, we can see that there was no significant difference between the U.S. and Hong Kong (ninth), and Japan (fourth) in cognitive domains knowing, applying, and reasoning questions. There was also no significant difference between the U.S. and Singapore (second) in cognitive domains knowing and applying (all of these nations ranked higher than the U.S.).

By digging a little deeper I was able to conclude that the U.S. is lagging behind in some content and cognitive domains, but not all of them. I was able show in which specific content and cognitive domains the U.S. is lagging and which ones we are in line with some of the leading nations. With this information, teachers, administrators, and policy makers can focus their efforts on the areas in most need. Relative rank of nation on international tests does not give the specific areas in which nations need to focus their efforts. Thus, the current study was able to add to the large body of knowledge pertaining to national educational achievement.

Recommendations for Teachers, Administrators, and Policy Makers

Disaggregating the TIMSS data has helped me to see the areas that the U.S. struggling in comparison to other top ranking nations. Teachers, administrators, and policy makers can use the current study to help them to focus on areas that need to be better understood by students. Specifically, policy makers need to know that, as a nation, our students are not making the growth necessary as they move from fourth to eighth grade. U.S. students are falling behind by the time they get to the eighth grade. According to the findings related to question 1: growth of average scale scores over time, the U.S. is lagging behind top scoring nations in a number of areas (i.e., life science and physical science for the fourth grade; chemistry, and physics in science for the eighth grade; all content areas for mathematics, especially geometry, for the fourth grade; all content areas for mathematics, especially geometry, for the eighth grade).

Teachers could use the current study to guide their daily lessons. Administrators could use the current study to guide the focus of their schools and districts. Policy makers could use the current study to guide the laws and allocation of funds based on content and cognitive domains.

Call for Further Research

There are many aspect of the current study that can be replicated for other international tests. One such international test is the Program for International Student Assessment (PISA), which also reports the rank of the participating nations (OECD, 2010; Olsen, & Lie, 2011). By replicating this study for other international tests, additional areas of need can be uncovered.

The TIMSS designers should not only publish the significant difference between participating countries and the international average, but they should also publish the significant difference between each participating country and all other participating countries. In the current study, I was unable to use all of the test item statistics because many of the test items were not released to the public since they will be used in subsequent iterations of the TIMSS. If policy makers and news organizations had access to tables that reported what countries were significantly different from all other countries, then they would be able to report a more accurate rank of each nation. The actual rank would be created by reporting all nations that are not significantly different from the nation in question into one position of rank. For example, if the U.S. is not significantly different from five nations above them on an international test, but are significantly different from three nations above them on an international test, then the U.S. rank would move from ninth to fourth. Ranking fourth would be a better representation of where the U.S. stands against other nations.

Performing *t*-tests on all nations that rank higher than the U.S. on the TIMSS could help uncover the actual rank, as mentioned above, of the U.S. on the TIMSS. The study could be performed using the same procedures and methods of analysis found in the current study and would give a better representation of the current standing of the U.S. on international tests.

Another way in which the current study could be replicated is by comparing the U.S. scores with scores of states that participated in the TIMSS. Minnesota and Massachusetts both participated in the 2011 TIMSS and they ranked in the top ten in multiple areas. By finding out in which areas that Minnesota and Massachusetts excelled, policymakers would have a local education system from which examples of success could be drawn.

My aim for the current study was to present a clearer picture of the data found in the TIMSS. I believe that by answering these questions and pursuing these conclusions, I have provided a clearer picture of the meaning of the U.S. rank on the TIMSS. As we implement policies based on accurate analysis, we can most effectively improve the U.S. education system.

References

Amrein, A.L. & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18), 1-74.

Arenson, K. W. (2004, December 15). Math and science tests find 4th and 8th graders in U.S. still lag many peers. *The New York Times*. Retrieved from http://www.nytimes.com

Armario, C. (2010, December 7). Wake up call: U.S. students trail global leaders. *NBC News.* Retrieved from http://www.msnbc.msn.com

Asimov, N. (2000, December 6). No gain by U.S. students on international exam: Math, science scores stay only above average. *San Francisco Chronicle*. Retrieved from http://www.sfgate.com

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1997). *Science achievement in the middle school years: IEA's IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beaton, A. E., Mullis, I. V.S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Blomeke, S., Suhl, U., Kaiser, G., & Dohrmann, M. (2012). Family background, entry selectivity and opportunities to learn: What matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education, 28,* 44-55.

Bracey, G. W. (2007). U.S. performance in international comparisons: An overview. *Principal Leadership (Middle Sch Ed), 7,* 66-69.

Burstein, L. (Ed.). (1993). *The IEA study of mathematics III: Student growth and classroom processes*. Oxford, United Kingdom: Pergamon Press.

Cardno, C. A. (2012). Improving U.S. education and surpassing Shanghai. *Education Week*, *31*(31), 12.

Carnoy, M., & Rothstein, R. (n.d.). *What do international tests really show about U.S. student performance?* Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/publication/us-student-performance-testing/

Chen, Y., Gorin, J. S., Thompson, M. S., & Tatsuoka, K. K. (2008). Cross-cultural validity of the TIMSS-199 mathematics test: Verification of a cognitive model. *International Journal of Testing, 8(3),* 251-271.

Husén, Torsten. (Ed). (1967). *International study of achievement in mathematics*. New York, NY: John Wiley.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Martin, M.O., Mullis, I.V.S., & Foy, P. (2008). *TIMSS 2007 international science report: findings from IEA's Third International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., Foy, P., & Stanco, G.M. (2012). *TIMSS 2011 international results in science.* Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 international science report: findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, J. S., Garden, R. A., & O'Connor, K. M. (2000). *TIMSS 1999 international science report: Findings from the IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA, International Study Center Lynch School of Education Boston College.

McKnight, C. C., Crosswhite, J. F., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective. A national report on the Second International Mathematics Study* (Report No. ISBN-0-87563-298-X). Champaign, IL: Stipes Publishing.

Mullis, I. V. S., (2012). Surpassing Shanghai: An agenda for American education built on the world's leading systems. *Comparative Education Review*, *56*(4), 719-729.

Mullis, I. V. S., Martin, M. O. (2006). *TIMSS in perspective: Lessons learned from IEA's four decades of international mathematics assessments.* Amesterdam, Netherlands: International Association for the Evaluation of Educational Achievement.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the primary school years: IEA's third international mathematics and science study.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics achievement in the final year of secondary school: IEA's third international mathematics and science study*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. E. (2000). *Gender differences in achievement.* Chestnut Hill, MA, International Study Center Lynch School of Education Boston College.

Mullis, I.V.S., Martin, M.O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth Ggrades.* Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth grades.* Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, J. S., & Smith, T. A. (2000). *TIMSS 1999 international mathematics report: Findings from the IEA's repeat of the third international mathematics and science study at the eighth grade.* Chestnut Hill, MA, International Study Center Lynch School of Education Boston College.

Mullis, I. V. S., Martin, M. O., Graham, J. R., O'Sullivan, C. Y., & Preushoff, C. (2009). *TIMSS 2011 assessment frameworks*. Amesterdam, Netherlands: International Association for the Evaluation of Educational Achievement.

Nagesh, G. (2010). Obama: US students' performance in science, math is unacceptable. Retrieved from http://thehill.com/blogs/hillicon-valley/technology/124789-obama-us-students-performance-in-science-and-math-is-qunacceptableq

Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in Gender Differences in Mathematics and Science (TIMSS 1995-2003). *Studies In Educational Evaluation*, *34*(2), 56-72.

OECD. (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics, and science (Vol. 1)*. Paris, France: OECD Publishing.

Olsen, R. V. & Lie, S. (2011). Profiles of Students' Interest in Science Issues around the World: Analysis of Data from PISA 2006. *International Journal of Science Education, 33*(1), 97-120.

Pelgrum, W.J., Eggen, T., & Plomp, T. (1986). *Second International Mathematics Study: The implemented and attained mathematics curriculum. A comparison of eighteen countries*. Washington, DC: Center for Education Statistics.

Robelen, E. W. (2013). New Global Results Spark Questions on Finland's Standing. *Education Week*, *32*(15), 8.

Robelen, E. W., & Gewertz, C. (2013). U.S. Students Exceed International Average, But Lag Some Asian Nations in Math, Science. *Education Week*, *32*(15), 8-9.

Scott, E. (2004). *Comparing the NAEP, TIMSS, and PISA in mathematics and science.* Washington, D.C.: National Center for Education Statistics.

Shen, C. (2005). How American middle schools differ from schools of five Asian countries: Based on cross-national data from TIMSS 1999. *Educational Research & Evaluation, 11*(2), 179-199.

Suter, L. E. (2000). Is student achievement immutable? Evidence from international studies on schooling and student achievement. *Review of Educational Research, 70,* 529-545.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901-926.

Tienken, C. H. (2013). TIMSS implications for U.S. education. *AASA Journal of Scholarship and Practice, 9,* 3-9.

U.S. Department of Education (2010, December 7). *Secretary Arne Duncan's remarks at OECD's release of the Program for International Student Assessment (PISA) 2009 Results*. Retrieved from http://www.ed.gov/news/speeches/secretary-arne-duncans-remarks-oecds-release-program-international-student-assessment

United States. National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform : a report to the Nation and the Secretary of Education, United States Department of Education*. Washington, D.C.: Superintendent of Documents, Government Printing Office.

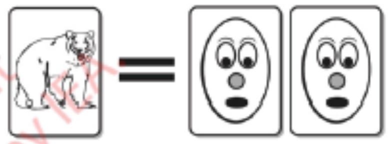
Ward, H., & Stewart, W. (2012). International rankings deliver mixed messages. Retrieved from: http://www.tes.co.uk/article.aspx?storycode=6309491

Appendix A

Example of Released Test Item

ID: M031346	Mathematics Grade 4	Block_Seq: /

The town fair had a booth where people could trade cards.



1 animal card is worth 2 cartoon cards.

2 animal cards are worth 3 sports cards.

Some children went to the booth to trade cards.

Questions for Trading Cards begin on the next page.

MP31346

Content Domain

Number

Topic Area

Whole Numbers

Cognitive Domain

Maximum Points

Key

See scoring guide

Continued on next page.

Trading Animal Cards

A. Becky had 5 animal cards to trade for cartoon cards. How many cartoon cards would she get?

Answer: ____________ cartoon cards

B. Jim had 8 animal cards to trade for sports cards. How many sports cards would he get?

Answer: ____________ sports cards

C. Katrina had 6 animal cards. She wanted to trade them for as many cards as possible.

How many cartoon cards would she get? ____________

How many sports cards would she get? ____________

Should she trade for cartoon cards or trade for sport cards?

Answer: ________________________

Content Domain

Number

Topic Area

Whole Numbers

Cognitive Domain

Applying

Maximum Points

1

Key

See scoring guide

M031346

Questions for Trading Cards continue.

Continued on next page.

Code	Response	Item: M031346A
	Correct Response	
10	10	
	Incorrect Response	
79	Incorrect (including crossed out, erased, stray marks, illegible, or off task)	
	Nonresponse	
99	Blank	

Appendix B

Example of Released Test Item Statistics

M01_01A (M031346A): Trading Cards_Cartoon cards
Constructed Response (1 Point)

Country	Percent Correct	
Hong Kong SAR	88 (1.8)	▲
Singapore	86 (1.5)	▲
Chinese Taipei	84 (1.8)	▲
Russian Federation	82 (1.8)	▲
Denmark	82 (1.7)	▲
Lithuania	82 (2.1)	▲
Finland	81 (1.9)	▲
Japan	79 (1.5)	▲
Portugal	78 (1.9)	▲
Sweden	77 (1.9)	▲
Netherlands	77 (2.3)	▲
Germany	77 (2.3)	▲
Korea, Rep. of	75 (1.8)	▲
Czech Republic	75 (2.2)	▲
Austria	74 (1.8)	▲
England	74 (2.2)	▲
Ireland	73 (2.7)	▲
Northern Ireland	72 (2.2)	▲
Slovak Republic	72 (2.3)	▲
Thailand	72 (2.5)	▲
United States	72 (1.2)	▲
Poland	71 (2.1)	▲
Belgium (Flemish)	69 (2.4)	▲
Slovenia	68 (2.7)	▲
Serbia	68 (2.4)	▲
Australia	66 (2.1)	▲
Croatia	66 (2.3)	▲
Norway	66 (2.4)	
Hungary	65 (2.4)	
Italy	65 (2.0)	
Spain	63 (2.5)	
Romania	63 (2.9)	
New Zealand	62 (1.9)	

Continued on next page.

International Avg.	62 (0.3)	
Georgia	59 (2.6)	
Malta	56 (2.1)	▼
Kazakhstan	56 (2.8)	▼
Chile	53 (1.9)	▼
Azerbaijan	53 (2.5)	▼
Armenia	50 (2.2)	▼
Turkey	42 (1.8)	▼
United Arab Emirates	41 (1.3)	▼
Iran, Islamic Rep. of	39 (1.9)	▼
Bahrain	38 (2.2)	▼
Qatar	32 (2.2)	▼
Tunisia	31 (2.5)	▼
Saudi Arabia	29 (2.5)	▼
Oman	24 (1.6)	▼
Morocco	24 (2.2)	▼
Kuwait	14 (1.6)	▼
Yemen	11 (1.6)	▼
Sixth Grade Participants		
Honduras	44 (3.1)	▼
Botswana	32 (2.0)	▼
Yemen	20 (2.5)	▼
Benchmarking Participants		
Quebec, Canada	77 (2.1)	▲
North Carolina, US	72 (3.5)	▲
Ontario, Canada	72 (2.3)	▲
Alberta, Canada	68 (2.5)	▲
Florida, US	64 (2.9)	
Dubai, UAE	52 (2.1)	▼
Abu Dhabi, UAE	35 (2.3)	▼

▲ Percent significantly higher than international average
▼ Percent significantly lower than international average

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

SOURCE: IEA's Progress in International Reading Literacy Study – PIRLS 2011

Appendix C

2011 TIMSS Released Items Used in the Current Study

Eighth-Grade Mathematics			Eighth-Grade Science		
Knowing	**Applying**	**Reasoning**	**Knowing**	**Applying**	**Reasoning**
m01 01	m01 08	m01 02	s01 01	s01 02	s01 04
m01 07	m02 03	m01 03	s01 06	s01 03	s01 11
m01 09	m02 04	m01 04a	s01 10	s01 05	s02 03
m02 01	m02 06	m01 04b	s01 12	s01 07	s02 05
m02 02	m02 08	m01 04c	s01 13	s01 08	s02 07
m02 05	m02 11	m01 05	s01 14	s01 09	s02 09
m02 07	m03 02	m01 06	s02 04	s02 01	s02 11c
m03 01	m03 08	m02 09	s02 08	s02 02	s02 13
m03 04	m03 10	m02 10	s02 10	s02 06	s03 01
m03 05	m03 11	m02 12	s02 11a	s02 14	s03 09
m03 06	m03 12	m02 13	s02 11b	s03 03	s03 12b
m03 07	m03 15	m02 14a	s02 12	s03 05a	s03 12c
m03 13	m03 16	m02 14b	s03 02	s03 05b	s05 10b
m05 01	m03 17	m03 03	s03 04	s03 07	s05 11a
m05 04	m05 03	m03 09	s03 06	s03 14	s05 11b
m05 06	m05 05	m03 14	s03 08	s05 01	s05 11c
m05 11	m05 08	m05 02	s03 10	s05 05	s06 06
m05 12	m05 09	m05 07	s03 11	s05 06	s06 08
m05 14	m05 13	m05 10	s03 12a	s05 07	s06 09
m06 02	m06 01	m06 04	s03 13	s05 08	s07 02
m06 05a	m06 03	m06 05b	s05 02	s05 10a	s07 06
m06 06	m06 08	m06 05c	s05 03a	s06 02	s07 08
m06 07	m06 10a	m07 10	s05 03b	s06 03	
m06 09	m06 10b	m07 12	s05 03z	s06 07	

Eighth-Grade Mathematics Continued			Eighth-Grade Science Continued		
Knowing	**Applying**	**Reasoning**	**Knowing**	**Applying**	**Reasoning**
m06 11	m06 10z		s05 04	s06 10	
m06 12a	m06 12c		s05 09	s06 12a	
m06 12b	m07 01		s06 01	s06 12b	
m07 02	m07 06		s06 04	s06 12c	
m07 03	m07 07		s06 05	s06 12d	
m07 04	m07 09		s06 11	s06 12e	
m07 05	m07 11		s06 15	s06 12z	
m07 08	m07 13b		s07 01	s06 13	
m07 13a	m07 13c		s07 04	s06 14	
			s07 07	s06 17	
			s07 11	s07 03	
				s07 05	
				s07 09	
				s07 10	
				s07 12a	
				s07 12b	
				s07 12z	
				s07 13	

Fourth-grade Mathematics			Fourth-grade Science		
Knowing	**Applying**	**Reasoning**	**Knowing**	**Applying**	**Reasoning**
m01 06	m01 01 a	m01 01 b	s01 01	s01 02	s01 05
m01 07	m01 05	m01 01 c	s01 07 a	s01 03	s01 10
m02 02	m02 01	m01 02	s01 07 b	s01 04	s02 09
m02 05	m02 06	m01 03	s01 07 z	s01 06	s02 10
m02 07 a	m02 07 b	m01 08	s01 09	s01 08	s03 06
m02 09	m02 08	m02 03	s02 01	s01 11 a	s03 10
m02 10	m03 02	m02 04	s02 05	s01 11 b	s05 04
m03 01	m03 03	m02 11	s02 06	s01 11 z	s05 10
m03 04	m03 09	m03 12	s02 11	s02 02	s06 08
m03 05	m03 10	m03 14	s02 12	s02 03 b	s06 09
m03 06	m03 11	m05 02	s03 01	s02 03 c	s07 02
m03 07	m05 03	m06 06 b	s03 02	s02 03 d	
m03 08	m05 05	m06 10	s03 03	s02 03 e	
m03 13	m05 06	m06 12	s03 04	s02 03 z	
m05 01	m05 08	m07 06 c	s03 11	s02 04	
m05 07	m05 09		s05 01	s02 07	
m05 10	m05 12		s05 02	s02 08	
m05 11	m06 01		s05 03 a	s03 05	
m06 02	m06 06 a		s05 03 b	s03 07	
m06 03	m06 07 b		s05 03 z	s03 08	
m06 04	m06 08		s05 07	s03 09	
m06 05	m07 02		s05 08	s03 12	
m06 07 a	m07 03		s06 02	s03 13 a	
m06 09	m07 06 b		s06 05	s03 13 b	

Fourth Grade Mathematics Continued			Fourth Grade Science Continued		
Knowing	**Applying**	**Reasoning**	**Knowing**	**Applying**	**Reasoning**
m06 11	m07 07		s06 07	s05 05	
m07 01	m07 08		s06 10	s05 09	
m07 04	m07 09		s06 12	s05 11	
m07 05	m07 10		s06 13	s06 01	
m07 11	m07 12		s07 01	s06 03	
			s07 06	s06 04	
			s07 07	s06 06 a	
			s07 09	s06 06 b	
			s07 10 a	s06 11	
			s07 10 b	s07 03	
			s07 10 z	s07 04	
			s07 11	s07 05	
				s07 08	