
by

Christiana Huss

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Biochemistry & Biophysics
(Honors Scholar)

Honors Baccalaureate of Arts in Spanish
(Honors Scholar)

Honors Baccalaureate of Arts in International Studies
(Honors Scholar)

Presented June 2, 2017
Commencement June 2017
AN ABSTRACT OF THE THESIS OF


Abstract approved:_____________________________________________

SueAnn Bottoms

The purpose of this meta-synthesis was to investigate why students from the U.S. receive the scores they do on the international TIMSS and PISA science exams. This work sought to shed light on the perceived disparity between science performance of U.S. students and students from other countries. To do so, twenty-two papers that offered possible explanations for U.S. science performance in an international context were reviewed and synthesized. One important finding was that several variables had different effects on science achievement in different countries; for the most part, there were no trends that indicated high-performing nations engaged in certain practices that low-performing nations did not. As such, in order to work towards better science education and achievement, these findings suggest that the U.S. might more effectively shape policy by modeling high-performing schools within the U.S., rather than by modeling other high-performing countries.

Key Words: TIMSS, PISA, science education, science performance/achievement

Corresponding e-mail address: hussc@oregonstate.edu

by
Christiana Huss

A THESIS

submitted to
Oregon State University
Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Biochemistry & Biophysics
(Honors Scholar)

Honors Baccalaureate of Arts in Spanish
(Honors Scholar)

Honors Baccalaureate of Arts in International Studies
(Honors Scholar)

Presented June 2, 2017
Commencement June 2017
Honors Baccalaureate of Science in Biochemistry & Biophysics and Honors Baccalaureate of Arts in Spanish and Honors Baccalaureate of Arts in International Studies presented on June 2, 2015.

APPROVED:

_____________________________________________________________________
SueAnn Bottoms, Mentor, representing the College of Education

_____________________________________________________________________
Matt Nyman, Committee Member, representing the College of Education

_____________________________________________________________________
Rebekah Lancelin, Committee Member, representing International Studies Degree Program

_____________________________________________________________________
Toni Doolen, Dean, Oregon State University Honors College

I understand that my project will become part of the permanent collection of Oregon State University, Honors College. My signature below authorizes release of my project to any reader upon request.

_____________________________________________________________________
Christiana Huss, Author
ACKNOWLEDGEMENTS

I want to extend my sincere gratitude to my mentor, Dr. SueAnn Bottoms. She has aided me immeasurably through this process, from helping me formulate this project to teaching me how to write a paper in a field different than my own. Without her expertise and generous time, this thesis would not be possible.

I would also like to thank my committee members, Dr. Matt Nyman and Rebekah Lancelin. Their support and time was vital for my success with this thesis.
Introduction

Comparing the United States’ performance in science to nations across the globe, mainstream media relentlessly decry the substandard education system here and students’ seemingly pitiable science achievement. With titles varying from “U.S. Students Fail to Impress in Math and Science” to “American Schools Are Failing Miserably at Science Education,” the media bombards us with the notion that the current state of science education in this country is in dire straits (Metcalfe, 2012; Williams, 2017). Such proclamations encourage questions surrounding the validity of these claims. How do students from the U.S. perform in science when compared to other nations? How are these differences verified? If U.S. students’ science achievement is significantly different than that of students from other countries, what factors contribute to this disparity?

Test scores indicate that students from many developed countries perform markedly better than U.S. students when taking the same tests, specifically the international exams the Program for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), which test students around the world on multiple subjects including science. Since the conception of these exams, youth in the U.S. have consistently performed below the international average on PISA science and at or below the international average on the composite TIMSS (Loveless, 2017). These numbers seem to indicate that U.S. students are underperforming in the sciences. However, they do not help us understand what accounts for these scores.

In this meta-synthesis, my goal is to investigate the science achievement of U.S. youth, as measured by TIMSS and PISA, in an international context. I want to
understand what might account for the differentiation in scores of U.S. students from those in other countries that consistently outperform U.S. students in science. Looking at nations that receive scores higher than those in the U.S. may have the potential to shed light onto ways science education could improve, and thus may have implications for education policy in the U.S. Similarly, factors and patterns noticed in data from countries that perform worse than the U.S. may provide information about methods and practices that are less useful in providing effective science education. By compiling research focused on reasons for the poor or mediocre science test scores of the U.S., I aim to address the following questions:

- What factors has previous research identified as contributing to the disparity between K-12 science performance in the U.S. and other nations?
- How much do these factors help explain the differences in test scores?

**Background**

**TIMSS and PISA structure.** The TIMMS exam has been administered every four years since 1995 by the International Association for the Evaluation of Educational Achievement, an organization that has been overseeing international tests of all subject areas for more than half a century (Nixon & Barth, 2014). While this Association is formally based in The Netherlands, much of the exam has been created in the U.S. at Boston College (Bracey, 2009). The first round of TIMSS included fourth graders, eighth graders, and students at the end of high school. This final group, however, was challenging to define because compulsory education varies significantly across countries, and has thus been excluded from all subsequent TIMSS exams (Koretz, 2009). The fourth grade has been defined, to account for differences in grade structure, as the fourth
year of formal schooling, provided that student mean age at this point is 9.5 years old, and eighth grade has been defined as the eighth year of formal schooling, provided that student mean age is 13.5 years old. Each country must formally define these target grades and must continue with these definitions in subsequent test participation (Joncas, 2007).

Sampling methods for TIMSS have required countries first to define their desired target population, from which student participants can be chosen. Students have been excluded from each country’s desired target population as infrequently as possible. Either whole schools or individual students can be excluded from the target population, so long as, in all, no more than 5% of a nation’s students are excluded. Grounds for exclusion have included differences in the first language, geographic constraints, special-needs, and schools with “curriculum or structure…different from the mainstream education system” (Joncas, 2007, p. 80).

While participating countries have been able to elect to test their students in fourth grade, at eighth grade, or at both grade levels, sampling methods have remained consistent worldwide with few exceptions. A two-stage stratified cluster design of probability proportional to size has been used for sampling. In the first step, schools have been randomly selected within each country, with larger schools more likely to be chosen than smaller schools. Then, classrooms within each school have been chosen, with classrooms with more students being more likely to be selected. Whole classes have participated. Russia and Singapore have each added a third level to their sampling; in Russia, regions have been sampled before schools, and in Singapore, individual students have been sampled within classes. All countries have been required to test 400 students
from a pool of at least 150 schools and 4,000 students (Joncas, 2007).

TIMSS has been primarily composed of multiple-choice, “fact-oriented” questions, making it widely considered a test of math and science curriculum (Bracey, 2009). More specifically, TIMSS has been designed with three cognitive domains for both subjects, namely knowing (identifying facts and methods), applying (problem solving), and reasoning (finding solutions to complex problems) (Nixon & Barth, 2014). In science, the “knowing” domain has asked students to describe, define, illustrate, recall, and utilize tools. The “applying” domain has asked students to find solutions, make comparisons, classify information, use models, and give explanations. The “reasoning” domain has asked students to extend their knowledge on a deeper level by tasks such as synthesizing, making hypotheses, analyzing data, and drawing conclusions. TIMSS has been divided so that 40% is knowing, 40% is applying, and 20% is reasoning. In addition to the exam, schools and students have been administered a post-test questionnaire to obtain data on students and their schools, from demographic information to student opinions (Nixon & Barth, 2014).

PISA has been administered to 15-year-olds worldwide triennially since the first exam was administered in 2000. Three subject areas – math, reading, and science – have been tested each year, but one subject has been the focus of each exam. The Organization for Economic Cooperation and Development (OECD), based out of Paris, has administered the test to approximately 30 OECD member countries, and to a comparable number of partner nations each year. Nations must have 5,000 students to participate (Bracey, 2009). While these have included both part-time and full-time students, the test has excluded homeschoolers, youth in the workforce, and youth abroad
PISA test participant selection has been through a pool that countries created from eligible students that ranged from 15 years and three months old to 16 years and two months old at the time the test was administered. Students have then been chosen based on a two-stage stratified design, very similar to that used to choose TIMSS participants. Schools with 15-year-old students enrolled have been included in a comprehensive national list, and 150 schools from each nation have been randomly chosen to participate based on probability proportional to measure of size, with measure of size defined as the number of 15-year-olds enrolled at the school. In the second stage, a target cluster size has been set; 35 students have been chosen randomly from each school, except for schools with fewer eligible students, in which case all have been included. These smaller schools must have had at least 20 students to be eligible. In all, a pool of 4,500 students from each country has been created for each test (or the full 15-year-old population from participating schools in countries with fewer students). Grounds for exclusion, like TIMSS, could be at the school or student level. These have included special needs (mental or physical) or differences in first language. For a country to participate, less than 5% of students could be excluded. Like with TIMSS, Russia has included a third-level of regional selection for PISA. However, Singapore has added no additional method of sample selection (Bertrand & Schleicher, 2014).

Various question types have been included in PISA science: in 2000, the exam was 45% multiple choice, 45% self-generated answers, and 10% limited range answers. The exam’s primary aim has been to test students’ ability to apply their knowledge and skills in a real-world context, as opposed to testing memorization or content knowledge
The OECD explains that they have designed the science test to measure “the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity” (Beese & Liang, 2010, p. 267). The exam has been broken into domains – for example, general science knowledge and scientific competency. Within these domains, questions have been designed based on various content categories, including physical systems, earth/space systems, living systems, and technology systems (Beese & Liang, 2010). Each exercise has been scored from 0 to 800 (Moore, 2002).

Thus, while these tests have been designed similarly in some regards, they are clearly not identical. Like with TIMSS, both students and school principals have been required to complete a questionnaire following PISA testing (Moore, 2002). A notable difference is that TIMSS has tested entire classrooms within a school, while PISA has tested random students within a school. Furthermore, the grounds for exclusion have been more comprehensive for TIMSS than for PISA, meaning that schools could be excluded from participating in TIMSS based on purported curricular differences, while this has not been grounds for exclusion in PISA.

**Test results.** U.S. student performance on TIMSS science has remained relatively constant since 1995, though there have been slight variations in average test scores. The average score for fourth graders in 2015, 546, was statistically higher than fourth graders’ scores in 2003 and 2007. However, it was not statistically different than average scores from other test years. For eighth graders, scores have improved from 1995 to 2015 (from 513 to 530), yet this has not been a consistent increase. The average
score in 2015, 530, was statistically higher than scores in 1995, 1999, and 2007, but was
not higher than 2003 or 2011 scores (Provasnik, Malley, Stephens, Landeros, Perkins, &
Tang, 2016). Thus, we see that there has been no definitive, long-term trend in U.S.
science achievement for either fourth or eighth graders, as measured by TIMSS.

In official statistics, an international average score is not reported. Rather, the
TIMSS benchmark of 500 points has been used each year as a point of comparison,
meaning that U.S. students have always been above the benchmark for science in both
fourth and eighth grades. In the most recent exam, the highest scoring fourth-grade
country (Singapore) received a score of 590, while the lowest scorer (Kuwait) received
339 (Provasnik, et al., 2016). The highest score has remained consistently in this range
and has been either Singapore or Korea, while the lowest score has varied significantly,
from 416 to 197 (Nixon & Barth, 2014).

Though average scores have remained relatively constant, in the U.S.
international ranking has been consistently decreasing (Nixon & Barth, 2014). In 1995,
U.S. fourth graders ranked third in the world, sixth in 2003, eighth in 2007, seventh in
2011, and eighth in 2015 (Nixon & Barth, 2014; Provasnik et al., 2016). It is important
to note, here, that the participant pools have not been constant; 17 countries participated
in 1995, versus 53 in 2017. Also of note is that some states in the U.S. have elected to
represent themselves in the exam; thus their scores are reported separately from that of
the U.S. and factor into the international rankings (Nixon & Barth, 2014). Similar
patterns have been seen among U.S. eighth graders, both regarding score ranges and in
changes in international ranking (Nixon & Barth, 2014; Provasnik et al., 2016).

On the other hand, students from the U.S. have consistently received scores either
at or below the OECD average on science PISA; the OECD does report an annual average, as opposed to using an invariable benchmark. U.S. scores have fluctuated between 489 and 502 since the PISA science test was first administered in 2006, and the average has typically been around 500. However, the differences in these scores over time have not been statistically significant, meaning that while the U.S. has been either at or below the international average, U.S. students are neither improving nor worsening in PISA science performance (Loveless, 2017). The range of scores among other nations has been similar to that seen with TIMSS; for example, the top-scoring participant in 2015, Shanghai, received 580, while the lowest scorer that year, Peru, received 373 (OECD, 2012).

**Literature Review**

In this section, I will explain the methods used in this meta-synthesis, after which I will present my findings from 22 papers that have focused on potential reasons for differences in science performance on TIMSS or PISA. I designed my thesis using the guidelines outlined by Walsh and Downe (2005) in their paper detailing how to analyze qualitative research in a meta-synthesis.

**Acquisition of Studies**

I found the papers included in this meta-synthesis with various search tools, including Google Scholar, OSU Libraries 1Search, PubMed, and EBSCO Host. Queries included variations of the phrase “United States PISA science scores”, where “United States” could also be “U.S.” or “USA”, where “PISA” could also be “TIMSS”, and where “science test scores” could also be “science performance”, “science achievement”, “science results”, or “science”. I also ran queries with specific years of test cycles (e.g.
TIMSS science scores 2015). I read the abstracts of 77 studies, and 48 of these in full. I then analyzed each study according to the criteria for selection. Ultimately, I included 22 articles in this meta-synthesis, 12 primary studies and 10 reviews.

**Figure 1.** Literature search strategy. Adapted from Li, Grimshaw, Nielsen, Judd, Coyte, & Graham (2009).

**Criteria for selection.** To be included in the meta-synthesis, each study met the following criteria:

- Was published between 1995-2017
- Used PISA and/or TIMSS results as a measure of science achievement
- Explicitly considered U.S. science achievement in an international context (i.e. studies done with subpopulations of the U.S. were not included)
- Gave reasons or explanations for differences in science test scores

I have divided the literature review into eight subsections based on recurring topics found throughout these studies. Two of these topics (school-level factors and test criticism) are further broken into sub-categories because of the extensive focus placed on
these topics within the literature analyzed.

**Socioeconomic Status**

Multiple studies throughout the 1970s suggested significant connections between familial socioeconomic status (SES), national economic development, and student science achievement (Heyneman, 1976; Jencks, 1972; Nichols & Anderson, 1973). More recent studies and changes in socioeconomic environments have challenged these findings (Moore, 2002; Chiu, 2007; Kaya & Rice, 2010). Emerging from this early research came the Heyneman-Loxley (HL) Effect, which describes how national Gross Domestic Product (GDP) influences student achievement (Baker, Goesling, & LeTendre, 2002). Essentially, the theory indicates that national GDP is a predictor of familial SES, ultimately influencing student achievement and that national GDP influences the quality and quantity of school resources, which also influences student achievement (Baker et al., 2002).

Moore (2002) analyzed international performance in math, reading, and science in the 2000 PISA exam and noted that contrary to popular conception, a nation’s wealth did not correlate with test performance. Some of the top performing students were from the poorest countries, and students from some of the richest countries did very badly on the exam. Furthermore, national wealth has not been the only factor at play here; countries that spend less on education, too, have often outperformed countries with higher education budgets (Moore, 2002). We see this looking at countries such as Korea and Finland, which have performed well above the PISA international average but spend less than average, or the U.S., which is one of the top spenders and yet has performed below average. Nevertheless, Moore did not expound upon these observations: his conclusions
were drawn from looking at composite PISA scores, which include math, reading, and science. Thus, while such observations might suggest a tie between U.S. students’ science achievement and GDP, we must look explicitly at this connection.

Baker et al. (2002) did directly investigate how GDP and science achievement were related. Their analysis of TIMSS science revealed that the HL effect did not hold, indicating that GDP was no longer a strong predictor of science achievement via its relationship with familial background or school resources. In particular, national income was, as of 1994, not associated with “school-resource effects or the size of school effects”, (p. 302) factors which were previously shown to influence science achievement. Thus, though students participating in TIMSS come from countries with different levels of economic development, they are not from backgrounds that would statistically skew their test results. These findings were true for both math and science (Baker et al., 2002).

Though GDP could not predict familial SES or school resources across participating nations, throughout nations, students from families of lower SES scored lower on TIMMS science; this trend between familial SES and science scores was more significant than was that between school resources and science scores (Baker et al., 2002). Kaya and Rice (2010) also analyzed this connection, looking at familial SES and TIMSS 2003 test scores. They compared a smaller group of countries, including the U.S., Singapore, Japan, Australia, and Scotland. Familial SES had a positive correlation with science scores within all countries except Japan; science scores were not correlated with familial SES there. Milford, Ross, and Anderson (2010) found, comparing North and South American countries who participated in PISA, that in all countries, familial economic, social, and cultural status (ESCS) was positively correlated with science
literacy scores. However, this relationship was significantly stronger in some countries (Canada and the U.S.) than it was in others (Brazil and Chile).

It seems, then, that the relationship between familial SES and science achievement is more complex when looking on a smaller scale and when considering individual nations. Chiu (2007) suggested one reason for this disparity might arise because of differences in national wealth. She posited that familial SES is more pertinent for students in economically homogenous nations because students tend to associate more readily with others of the same familial SES. Thus, “greater income equality within a country might also encourage greater cooperation among students, resulting in higher overall academic performance” (p. 511). In all, these studies suggest that familial SES is a fairly reliable predictor, within nations, of science performance, but this is not always accurate, nor is it a sure way to predict science performance of two students from different countries.

Though differences in GDP were not accurate predictors of test scores via connections to familial SES or school resources, Baker et al.’s (2002) analysis did reveal that GDP and TIMSS science scores were generally positively correlated on an international scale. Thus, both familial SES and national GDP are related to science performance, though familial SES and national GDP appear to share a complex relationship. Chiu (2007) offered insight by further exploring the effects of GDP after accounting for past academic records, by analyzing if, and how many, remedial courses TIMSS participants had taken. Past achievement had shown to be significant on PISA performance; students enrolled in remedial courses scored from 31 to 34 points below their peers (Chiu, 2007). Chiu discovered that, after accounting for past performance,
students from countries with higher GDPs scored higher on science than students from countries of lower GDPs, generally speaking. Additionally, students from countries with greater income equality scored higher in science than students from countries with less income equality. These conclusions support that GDP is, in fact, positively related to science achievement, and they also reveal that a country’s overall wealth is not the sole factor influencing science achievement; distribution of wealth, too, is pertinent.

**Familial Structure**

Chiu (2007) investigated the effects of familial structure on PISA 2006 science by considering two hypotheses. The resource provider hypothesis states that the presence of additional family members such as parents can provide resources (knowledge, help with homework, books) that academically benefit their children. The other, the research dilution hypothesis, states that additional family members such as grandparents and siblings can reduce resources necessary for a child to succeed academically; in this case, the time and money parents spend on these extra family members reduces the focus on any one child/student. While the aforementioned family members are typically in those roles of providers or diluters, this is not always the case – for example, affluent, healthy, educated grandparents can be providers rather than diluters.

To understand the effects of familial structure, Chiu (2007) performed three-level hierarchical linear modeling using post-test questionnaires, and the results were consistent across 41 nations: after normalizing for past science performance and GDP, different familial structures consistently corresponded to trends in science achievement. Children with more parents, who came from a higher SES, and who were native-born performed 17 points higher than average. First-generation immigrant students received
scores 21 points lower than average, while this figure was eight points for second-generation students. Furthermore, students coming from homes where the language used at school was not the predominate language spoken at home averaged 17 points lower. Students living with grandparents and siblings received scores 14 points lower on average (Chiu, 2007).

Other effects proved to be less significant to science achievement. For example, students coming from homes with large quantities of books only scored one point higher than those without such resources (Chiu, 2007). Similarly, Beese and Liang (2010) discovered, from the 2006 PISA data, that familial SES and the number of books found in the home had a small positive correlation with science scores when comparing students within the U.S. However, the authors did not see this trend in Finland; SES was not a predictor of science scores there (Beese & Liang, 2010). Milford et al. (2010) found that the highest level of parental education and the amount of home possessions (such as books) could account for 6% of the variance in PISA science scores between nations, whereas other school-level factors could account for as much as 72% of this variance. Chiu (2007) also looked at “cultural communication” (p. 516) and found that students who regularly discussed cultural matters with their parents scored two points higher on science than those who didn’t.

Despite these findings, a given nation’s characteristics were also relevant to this discussion: “country properties were linked to academic achievement and moderated the links between family constructs and academic achievement” (Chiu, 2007, p. 516). While these trends of familial structure were consistent across nations, their effects were more significant for students from wealthier, more egalitarian nations, i.e. the difference in
scores between a non-immigrant student, with two parents, of a higher SES, and an immigrant student, with one parent, of a lower socioeconomic status was more significant for two students that fit these profiles in a wealthy country than for two students that fit these profiles in a poorer country. Thus, while a familial structure is influential on science achievement (as measured by PISA), it cannot be readily isolated from national characteristics when predicting science performance.

**Student Interest and Attitudes**

Several studies have explored the connections between students’ feelings toward science and their achievement in the subject (House, 2012; Kaya & Rice, 2010; Papanastasiou & Zembylas, 2004; Ainley & Ainley, 2011). Many authors referenced in this meta-synthesis used post-test questionnaires, administered with both TIMSS and PISA, to determine correlations between various factors and science achievement, and regarding this topic in particular. For example, in a comparative analysis of 2007 TIMSS scores and student attitudes between Japanese and U.S. students, House (2012) found that, in both countries, students that had faith in their abilities in science received higher scores than those who didn’t. In the U.S., students who believed themselves to be good at science (“I usually do well in science”) tended to receive high TIMSS science scores, while students who felt badly about their science abilities, both individually and in comparison to their peers (“I am just not good at science” or “Science is harder for me than for many of my classmates”), did poorly. Interestingly, students who indicated “I enjoy learning science” tended to receive lower scores in the U.S., something not seen in Japan. In Japan, students who indicated “I usually do well in science” or “I learn things quickly in science” received high scores. Students that compared themselves to their
peers with “Science is harder for me than for many of my classmates” also did worse on the exam. Thus, in Japan, there existed an association between quick learning and high science performance, a notion absent in the U.S. Also of note is that many U.S. students who performed well on these tests had negative feelings toward the discipline, while high-performing Japanese students did not (House, 2012).

Researchers have also studied the effect of overall classroom attitudes toward science. Kaya and Rice (2010) found that in Scotland and the U.S., students who were from classrooms with overall higher mean self-confidence in science outperformed those students from classrooms with lower means (comparing TIMSS science scores). Classroom mean self-confidence was unrelated to science scores in Japan, Singapore, and Australia. Taken together with House’s (2012) results, these findings suggest that U.S. students’ peer environments are certainly influential on their science achievement, and perhaps more so than for students from other countries.

Papanastasiou and Zembylas (2004) aimed to investigate the directionality of the relationship between attitudes towards science and science achievement: do positive attitudes toward science influence high achievement, or does high achievement result in positive attitudes? The authors compared high school seniors who were past TIMMS participants from Cyprus, Australia, and the U.S. They analyzed data obtained in a comprehensive study carried out by the International Association for the Evaluation of Educational Achievement (IEA), and focused on those questions relating to student attitudes, including how students believed science was important for themselves and those around them, their enjoyment of the subject, and their self-perceived science abilities. In Australia, high school seniors who fostered positive attitudes toward science
had not received higher TIMSS scores in elementary school. On the reverse, Australian high school seniors who had done well on the TIMSS exam did more commonly have a positive attitude toward science at the time of the survey, which suggests that young students who did well in science continued liking the subject at a later age, and even young students who had not excelled in science could later develop a positive outlook on the subject. Contrarily, in Cyprus, those high school seniors who had a positive attitude toward science had received high TIMSS scores. Students who had performed well on TIMSS science did not uniformly have a positive outlook toward science at the time of the survey, nor was their outlook uniformly negative. In the U.S., high school seniors with positive attitudes toward science had received high scores on TIMSS science. Interestingly, students who scored high on TIMSS often consequentially developed negative attitudes toward science (Papanastasiou & Zembylas, 2004). Similar to House’s (2012) findings, then, students from the U.S. who performed well on these international science tests commonly had negative outlooks toward science, a trend not seen in any other country analyzed here.

Technology

Researchers have investigated both the time spent with technology and the ways U.S. students use technology for their potential influence on academic achievement in science. Within the U.S., House (2012) found evidence that elementary-aged children who use a “collaborative computer experience” (p. 264) perform better in science, while high school students with access to virtual microscopes have both their achievement and their interest in science heightened. Thus, technology offers a clear potential for science education in the U.S.
Nevertheless, technology also poses many possible negative consequences, primarily regarding time spent on computer games, something more than 75% of adolescents from industrialized nations regularly did as of 2010 (Drummond & Sauer, 2014). The distracting nature of video games has been blamed for poor attention and quality of sleep among youth, as well as for compromising the ability of students to focus on their schoolwork, science and otherwise (Drummond & Sauer, 2014). The true effects of this factor are hard to monitor, as, like with student interest in science, it is difficult to determine which variable “comes first”: do students that spend excessive time gaming consequentially underperform academically, or do students who are low achievers in school consequentially resort to gaming? The understanding of video gaming is further clouded by recent research that suggests there are significant differences, both physiological and academic, between those who play video games alone versus with friends (Drummond & Sauer, 2014).

House (2012) looked at fourth-grade performance on 2007 TIMSS science, comparing the U.S. with Japan. In the U.S., there were five variables related to computer usage that had statistically significant effects on science scores. Those who performed well on TIMSS science indicated that they frequently used computers both at home and at other venues outside of school (an internet cafe, a library, a friend’s house), but not at school. Nevertheless, students who indicated that they frequently used computers for science schoolwork had relatively low TIMSS scores. Those who spent more time playing computer games or browsing the internet also received lower test scores. These results were similar among Japanese students, but not identical. Japanese students who performed well on science TIMSS indicated that they frequently used computers both at
home and at school, but not at venues outside of school. Also like U.S. students, those
who spent more time using computers for science schoolwork performed worse on
TIMSS. Students who spent more time playing computer games also did worse.

Drummond and Sauer (2014) also investigated the relationship between
technology and science, by seeing how video-gaming impacted science achievement as
measured by PISA 2009. More than 192,000 fifteen-year-old students from 22 OECD
countries were surveyed about their video-game usage. Participants were asked if they
gamed never/hardly ever, once/twice a month, once/twice a week, or daily. They found
that, contrary to popular belief, increased video-game usage was not associated with
decreased test scores in science; within countries and comparing countries, the variables
were not correlated. This lack of correlation was true taking account for both single-
player and multi-player gaming. What the authors did find was that increased video-
game usage was slightly correlated with decreases in reading proficiency; nevertheless, it
was not related to science performance (Drummond & Sauer, 2014). Taken together, the
results of these studies suggest that the relationship between technology and science
achievement is complex, and non-uniform internationally.

Cultural Factors

Ainley and Ainley (2011) hypothesized that “broader cultural” factors, like
“historical and traditional values,” may play a pivotal role in science interest and
achievement (p. 52). Simplifying this discussion to the ways that school or student
variables differ across nations provides us little without looking at the social landscape of
the countries themselves. However, this is certainly a challenging factor to include, as
one cannot comprehensively account for all the ways that a student’s culture may
contribute to their education.

Kaya and Rice (2010) contended that different cultural norms for parental involvement in a child’s science education might be significant. Many nations in East Asia perform consistently higher on science than does the U.S., and a possible explanation for this is that, as of 2010, approximately 70% of students there are placed in tutoring by their parents, while only 25% of U.S. students have this extra-curricular exposure to science. Similarly, an assessment following the 1995 TIMSS revealed that 41.9% of Japanese eighth-graders actively participated in out-of-school science lessons (Kaya & Rice, 2010). In 2003, eighth-graders from Japan who participated in independent learning activities scored significantly higher than their peers who didn’t (House, 2012). If extracurricular activities and education relating to science do, in fact, so significantly influence science performance, this may be one explanation for the relatively lower science achievement of U.S. students.

On a different note, one factor Darling-Hammond (2014) identified to explain disparities in science performance within the U.S. is racial and socioeconomic background. U.S. white and Asian students perform, on average, above the OECD average on PISA science, whereas African American and Latino students, on average, perform below this benchmark. Likewise, students from poor schools perform near the bottom of PISA in science while students from wealthier schools perform near the top. These trends are typically absent from more socioeconomically and racially homogenous nations (Darling-Hammond, 2014). Intimately related to this problem is school segregation. Schools in the U.S. with high populations of minority students tend to lack resources that their whiter counterparts have, including textbooks, small class sizes,
computers, and qualified teachers (Darling-Hammond, 2014). Because of this lack of resources, these schools frequently shorten both their school days and school years. Wealthier states in the U.S. spend about three times as much per student as poorer states (Darling-Hammond, 2014). The absence of such disparities related to race and wealth in other nations may, then, be a reason that they perform better than the U.S.

Furthermore, the U.S. has the highest rate of childhood poverty of all OECD countries, sitting at 22% as of 2014 (Darling-Hammond, 2014). By contrast, accounting for governmental programs such as food and housing assistance, most of the countries in the OECD have childhood poverty rates below 10%. Factoring similar U.S. governmental programs into our childhood poverty rate does not statistically change it. Growing up in poverty significantly impacts a child’s learning, as there are fewer opportunities within the home to begin learning on their own. By four-years-old, children from low-income families have a vocabulary one-third as developed as those from middle-income families, and similar educational disparities are seen throughout first grade across all subjects (Darling-Hammond, 2014).

The role of teachers is also viewed differently in many countries than it in the U.S. For example, in Finland, schools and teachers make close connections with professionals (such as textbook manufacturers), closing the gap between those shaping and administering education (Moore, 2002). Teachers are also very close with their students – home visits are a common occurrence. Because of this, education is more personalized. Moore (2002) noted that overall, only 66 of the 4070 schools in Finland in 2002 were private, attesting to the government’s role in personalizing education. Teaching in Korea is similar to teaching in Finland. Though it may not be the most
economically rewarding profession, it is nevertheless regarded highly. Teachers are constantly expanding their knowledge through professional development and further education, and the profession is over-subscribed. As these characteristics are absent from the U.S., Moore (2002) cited them as possible culprits in low science achievement; U.S. students have teachers with different societal outlooks and different opportunities for training and growth.

School-level Factors

The factors relating to science performance considered up to this point have been primarily focused outside of the classroom. Here, I consider characteristics of schools that various studies (Beese & Liang, 2010; Milford et al., 2010; Baker, Fabrega, Galindo, & Mishook, 2004; Kaya & Rice, 2010; House, 2012) have suggested being pertinent to science performance. The majority of these findings came from PISA and TIMSS post-test questionnaires, wherein both students and school officials reported on the conditions of their schools.

School resources. Beese and Liang (2010) investigated the effects of various school resources on science performance in PISA, including shortage and quality of science teachers, instructional resources, and student backgrounds. They compared the U.S., Canada, and Finland, and found that school resource variables had a non-generalizable relationship with science test scores. The school resource factors they tested, including school type (private vs. public) and lab equipment, showed no relationship with PISA science scores among Finnish students, yet these factors were correlated with the performance of U.S. and Canadian students. In both countries, students from private schools and with more lab equipment outperformed their peers.
Accounting for differences in SES, however, students from public schools outperformed students from private and charter schools in the U.S. While there was a clear correlation, these variables proved to be statistically significant only for Canadian students, not for those from the U.S.; the trend was too small to be outside of statistical error. Looking at North and South American countries who participated in 2006 PISA, Milford et al. (2010) also found that school-level ESCS did not have a direct relationship with science scores. In Canada, higher level school ESCS was associated with lower test scores, while in Mexico, Uruguay, and Colombia, higher ESCS was associated with higher test scores. Furthermore, a higher quality of school resources was significantly associated with higher PISA science scores in Brazil and Argentina, but this was not seen in other countries tested (Canada, Chile, Colombia, Mexico, the U.S., and Uruguay). Considering these trends together, then, we see that the effects of school resources on science performance is neither straightforward within the U.S. nor internationally.

**Instructional time.** In 1994, Congress created the National Education Commission on Time and Learning (Baker et al., 2004) to investigate the state of instructional time in U.S. schools. The commission determined that students from many other countries – for example, France, Germany, and Japan (a frequent high performer in science) – received more than twice as much instructional time as U.S. students. Since then, this variable has been approached in various ways, from looking at the days in the school year to the hours dedicated to any particular subject. Regardless, this is a challenging variable to define, as it is no small task to determine how much time is spent on actual instruction, versus time spent preparing for a class or checking homework (Baker et al., 2004). Making connections between science class time and science
performance then inspires the question of which activities or aspects of a class should be considered class time.

Regardless, numerous researchers have investigated the potential connections between science performance and science instructional time. For example, Baker et al. (2004) analyzed three datasets: TIMMS 1999, PISA 2000, and The International Study of Civic Education (1999), a cross-national exam similar to TIMSS and PISA but testing students in their civics knowledge. They included these three subject tests for comparison purposes. Here, Baker et al. (2004) defined instructional time as the total weeks of instruction per school year, class periods for a given subject in a week, and minutes spent in each class period. Ultimately, they found that “there is no significant relationship, at the cross-national level, between achievement test scores and the amount of instructional time” (a finding true for both science and math) (p. 322). Nations who spent more time on science instruction did not uniformly perform better than nations who spent less time on science instruction (Baker et al., 2004).

Baker, et al. (2004) then looked at this relationship within nations. In some countries – but not all – there was a weak relationship seen for science instructional time and test scores, but the extent of this relationship varied significantly. In some countries, more time spent on science instruction was correlated with higher test scores, while in other countries, more time spent on science instruction was correlated with lower test scores. While there was a small positive correlation in the U.S., it was not statistically significant, ultimately leading to the conclusion that there is no correlation between instructional time and achievement in science. Similarly, Kaya and Rice (2010) found that within Japan, Australia, Singapore, Scotland, and the U.S., time spent in science
class each week and had no impact on TIMSS science scores.

Additional light may be shed on this matter by more closely analyzing Baker et al.’s (2004) findings. The authors further divided PISA participants by whether they were in the ninth or tenth grade. In several cases, results were inconsistent between ninth and tenth grade PISA participants from the same country. For example, there was no correlation between instructional time and PISA score among ninth grade students from the Czech Republic, while tenth grade Czech students showed a negative correlation. This, taken together with Kaya and Rice’s results (2010), which looked at fourth and eighth graders, may suggest a difference in the effects of science instructional time based on grade level. Nevertheless, as these ninth to tenth grade differences for PISA science were not seen within the U.S., and as the correlations with TIMSS scores were too small to be statistically significant, this may not be the case for our students. These findings together suggest that neither increasing nor decreasing science instructional time would provide a straightforward route to improve science performance.

**Instructional methods.** Beese & Liang (2010) further revealed that the type of science instruction – not just its time – may be crucial. Within the U.S. and Canada, the more time students spent conducting experiments within a lab setting, and the more they indicated on the post-test questionnaire that their teachers offered clear explanations, the worse they did on PISA. They did not see these correlations in Finland; Finnish students showed no relationship between lab time and test scores, nor between extent of teacher explanation and test scores.

Similarly, House (2012) looked at correlations between instructional strategies and TIMSS science scores. In the U.S., two variables were correlated with higher test
scores: students spending time memorizing facts, and students working through problems individually during class time. Four factors were correlated with lower test scores: students designing their own experiments, watching the teacher demonstrate a science experiment, recording their observations, and performing their own science experiments and investigations. Like with U.S. students, Japanese students who spent time memorizing facts and working through problems individually did well on TIMSS.

Another factor that showed a strong positive correlation with high test scores in Japan was students who indicated, “I write or give an explanation for something I am studying in science” (p. 269), a relationship not seen with the U.S. students. Designing science experiments and watching teacher demonstrations, like with U.S. students, was correlated with lower test scores among Japanese students.

Kaya and Rice (2010), too, found that science inquiry, including such experimental and lab-based activities, has an unclear effect on science achievement. In Japan, Australia, and Scotland, classrooms incorporating more science inquiry performed the same as classrooms who incorporated less. In Singapore, more science inquiry was significantly correlated with higher mean achievement. In the U.S., more science inquiry was significantly correlated with lower mean achievement. Thus again, while there is no clear trend internationally as to the relationship between science instruction type and science achievement, we see that, within the U.S., more memorization and less experimental science lead to higher test scores. It may be advantageous to further explore what constitutes laboratory activities in the U.S. and in those countries where this is associated with higher test scores.

Teachers. Kaya and Rice (2010) found that in all countries besides Japan
(Singapore, the U.S., Scotland, and Australia), years of teacher experience and the field the teacher studied were not correlated with TIMSS science scores. In Japan, each additional year of experience was associated with a 0.40 higher score. In the U.S. and Singapore, students who cited their teachers were supportive performed significantly higher than those who didn’t (11 points higher in the U.S. and 15 points higher in Singapore). Such correlation was not seen in the other countries included in their study. Reviewing teacher preparation for science in Japan as compared to the U.S. could thus potentially be beneficial.

Looking at North and South American countries who participated in the 2006 PISA, Milford et al. (2010) found that student-teacher ratio also influenced science scores. In the U.S., a one-point increase in student-teacher ratio decreased overall science PISA scores by 2.5 points. Conversely, the same increase in the student-teacher ratio in Mexico raised science PISA scores by 12 points. Student-teacher ratio was not correlated to test scores in any other countries tested.

Beese and Liang (2010) found that science teacher shortage and the ratio between full-time and part-time science teachers showed to have no relationship with PISA science scores among Finnish students. However, lower rates of science teacher shortage and more full-time science teachers in both Canada and the U.S. were associated with higher science scores. Interestingly, Moore (2002) compared science teacher preparation in the U.S. with that in Finland, as the latter has a model he considered to be effective. Teacher preparation is, according to Moore, more rigorous in Finland than in the U.S., with teachers required to have at least five years of university-level education to teach primary school and six years to teach secondary school. Furthermore, teachers of the
natural sciences are trained extensively in “practical experiments” (Moore, 2002). While Moore did not provide further detail to support these claims, Evagorou, Dillon, Viiri, and Albe (2015) compared science teacher preparation in multiple countries including Finland. They commended the Finnish system for teacher preparation being research-based, rather than teaching based; throughout college, future teachers are trained in research methods, and are given ample opportunities to perform research. Evagorou et al. (2015) also noted that, while future primary school teachers are educated in comprehensive teaching methods, future secondary school teachers belong to the department of their study area (i.e. a future chemistry teacher majors in chemistry); their education classes come in the form of a Master’s program following their undergraduate career. Thus, as the status of Finnish teachers is not as impactful on science achievement as it is on science achievement in the U.S., this may point to weaknesses in U.S. science teacher preparation, which then ultimately affect how well students master the subject.

**Science Curricula**

Few studies have concretely defined the relationship between science curricula and international test performance. For example, Milford et al. (2010) stated that “governance, curricula, instructional methods, approaches to testing, and accountability” are all factors which countries that score on the lower end of TIMSS and PISA “do not do so well” (p. 460); their claim ends there. Nevertheless, though we may not have quantitative connections between curriculum and science performance, curricular differences are certainly recognized as influential on U.S. students’ science achievement. A common fault critics find with U.S. science curricula is its sheer volume. Noting the decline in science performance between fourth and eighth grade following the
third TIMMS exam, Schmidt and McKnight (1998) sought to investigate science curricula as a possible culprit. The authors noted that “U.S. curricula, as an aggregate, consistently covered more topics than did the curricula of virtually all other TIMSS countries” (p. 1830), which they used to explain U.S. students’ ability to keep up with the youth from other nations in the fourth grade, but to fall behind by the eighth. More recently, Beese and Liang (2010) again suggested that this wide inclusion of topics inhibits U.S. teachers from educating students in the areas of scientific innovation and flexible thinking, and on how to effectively solve problems in a real-world scenario. By attempting to teach U.S. students so much material, they lack the problem-solving skills necessary to excel on tests like PISA. Bracey (2009) noted this when comparing the U.S. to Hong Kong, as in the latter, many schools focus on three principal subject areas (math, English, and Chinese). Though science is not one of these, he contended that the well-defined nature of their curriculum enables more efficient and effective teaching. The breadth and variety of subject matter included in U.S. curricula, on the other hand, does not allow for such efficiency and effectiveness. Koretz (2009) also postulated that students from countries where specialized secondary schools are commonly an option (such as Germany and the Netherlands) often get a curriculum that is predominately scientific or technical in nature; the U.S. lacks schools with such a focus.

Schmidt and McKnight (1998) noted the size of science textbooks in the U.S. – consistently larger those of other nations. This multitude of topics ensures that teachers spend less instructional time in each area and that U.S. students are not exposed to certain topics until much later than students in many other countries (e.g. what a child from Japan may learn in sixth grade, U.S. students learn in the eighth). In fact, Schmidt and
McKnight (1998) suggested that this was the primary culprit behind U.S. students’ decline in performance. In fourth grade in 1995, Korea was the only nation to receive a raw mean science score higher than the U.S. Looking at specific topic areas, however, revealed that U.S. students did the worst on the physical sciences and physics – subjects which the typical fourth-grade textbook in the U.S. does not cover in depth. There is both less textbook content and less instructional time devoted to this subject in the U.S.

U.S. science curricula is not merely large; critics have also deemed it ineffective. For example, in addition to the sheer content of the average U.S. science curriculum, Schmidt and McKnight (1998) condemned it for being “highly repetitive, lacking coherence, and providing little rigorous intellectual challenge” (p. 1830). While children from other countries usually begin learning science topics such as chemistry and physics between the fifth and eighth grades, children in the U.S. often spend these years reiterating elementary science topics. Beese and Liang (2010) also cited the incoherence of U.S. science curricula in that teachers are not only given overwhelmingly exhaustive curricula, but they are also given curricula that can differ drastically from state to state. It would be difficult, then, to group the results from students of two states on the same test when what they are learning may differ as significantly as differences between countries.

The science curriculum of Japan in particular has been commended, and Japanese students consistently receive among the highest scores on both TIMSS and PISA (House, 2012). Kaya and Rice (2010) praised the Japanese science curriculum, drawing attention to its homogenous and focused nature in contrast to the extreme variability within the U.S. They criticized curriculum in the U.S. because, as a result of its large scope, most schools provide shallow coverage of any given topic. House (2012) noted that in Japan,
seventh and eighth graders spend a significant portion of their instructional time on science, and this time is not solely devoted to introducing new material. Their curriculum is designed to help students develop a positive outlook toward science (something U.S. students severely lack), to understand natural phenomena, and to engage in hands-on observations and experiments. Teachers commonly include problem-solving exercises in the classroom. The elementary school science lesson typically follows a format of introducing new material by connecting it to previous knowledge, performing investigations, organizing data, and discussing the findings in relation to initial hypotheses. House (2012) contrasted this format to classrooms in the U.S., which often lack this interconnectedness; U.S. students are frequently taught concepts in isolation.

Another problem identified in U.S. science curricula is the tendency for teachers to teach to the test. The American Association for the Advancement of Science (Osborne, 2013) stated that it is common for U.S. science curricula to focus on “the ability to commit terms, algorithms, and generalizations to short-term memory”, a teaching style that “impedes the acquisition of understanding” (p. 267). In other words, it is not uncommon for science teachers to teach to the test; students feel obligated to commit a wide volume of facts and figures to memory, hindering their ability to synthesize this knowledge in any useful way. Similarly, Moore (2002) identified teaching memorization as highly problematic within the U.S. The PISA exam consists of problems that are meant to be relevant to the real world, and avoids questions of regurgitation, wherein students are asked to write a complex formula or identify a specific date, yet students in the U.S. are still commonly prepared for these tests with memorization exercises. He cited unnecessary educational expectations such as
“memorizing the lanthanides” as an example of the “types of task that create problems for many students even in richer countries” (p. 297).

While U.S. curricula and teachers may have adopted such a format of memorization-based coursework to help students on these international tests, Beese and Liang (2010) noted the unique format of teaching in Finland. Education reform there eliminated standardized testing, and their standards were reduced to a nationally-used, ten-page guide from which the curriculum is developed. Since then, Finland’s students have excelled on international science exams. Their science curriculum centers on problem-solving and the use of technology to do so, independent learning, and higher-order thinking (Beese & Liang, 2010). Though Finland does not administer standardized tests, these are the areas that the PISA science focuses on, and their students typically excel.

Finland might be an exception, and the U.S. the norm. Osborne (2013) observed that while various tests, including TIMSS and PISA, aim to encourage higher level thinking, most countries still focus their science education around knowledge rather than application. Internationally, countries performed better on questions that required recall, definition, and description than on questions that required scientific reasoning. Osborne (2013) referred to the authors of an investigation involving university students in China and the U.S. who concluded that “‘it seems that it is not what we teach, but rather how we teach, that makes a difference in student learning of higher-order abilities in science reasoning’” (p. 267). Thus, while students from Finnish classrooms wherein memorization and teaching to the test were avoided performed better on PISA science, and while U.S. curricula are scrutinized for focusing on memorization, such a format
seems to be associated with higher science performance in most countries.

**Test Criticism**

There has been widespread consensus among policy makers, academics, and test experts alike that we should take the results from TIMSS and PISA with caution (Cavanagh & Manzo, 2009; Viadero, 2008; Schneider, 2009; Bracey, 2009; Koretz, 2009; Darling-Hammond, 2014; Nixon & Barth, 2014; Beese & Liang, 2010). Along with differences in curricula and teaching practices, these results are impacted by factors such as societal and cultural norms, demographics, and politics. Cavanagh and Manzo (2009) argued that, because of this, there is no way to account for all the elements contributing to why some countries do better than others. From the Brookings Institution in Washington, Tom Loveless, director of Brown Center on Education Policy, criticized PISA reports for being “‘fatally flawed’”, as correlational data cannot accurately account for the intricacies of education and educational disparities (Viadero, 2008, p. 10). He urged policymakers not to respond to the information media presents regarding tests like TIMSS and PISA. Similarly, Schneider (2009) stated that “international assessments cannot generate a great deal of reliable policy device” (p. 69), as these tests are merely an indication of how U.S. students compare to other nations – not on how exactly U.S. schools need to, and could feasibly, improve.

**Variation among nations.** One criticism of comparing the results of these tests is that the countries participating are so inherently different that it is not meaningful to perform ranking comparisons. For example, Bracey (2009) contended that it is unequitable to compare the U.S., comprised of 50 semi-independent states and over 300 million people, with countries like Singapore, drastically more homogenous and less
Koretz (2009) echoed these sentiments, noting that the complex political and size differences between countries used to compute test averages make these values false comparators. State-wide differences are certainly another point of concern. Comparing average PISA scores of the U.S. with countries like South Korea, we see performance gaps similar to those between Mississippi and Massachusetts on the National Assessment of Educational Progress (NAEP), a standardized test within the US (Cavanagh & Manzo, 2009). As noted previously in discussing curriculum, this begs the question of the validity of comparing the U.S., as a whole, to other countries.

**Test structure.** Issues related to test structure are another source of criticism. PISA aims to create questions that are not solely knowledge-based, but that require students to apply what they’ve learned to the real world. Bracey (2009) argued that such a structure intrinsically favors affluent students, as they have greater opportunities for learning within the home. More affluent countries would then, he posited, be predisposed to do better. Koretz (2009) further faulted these international tests because while they aim to test students’ broad mastery of a variety of subjects, only a small, non-representative sampling of questions is included. These questions may be biased towards the knowledge and interests of those who create the test, which may favor students from certain countries and disadvantage others.

Translation is another point of interest. PISA attempts to maintain the integrity of each question by high-fidelity translations between languages. Bracey (2009) referenced language experts who contend that this not only creates awkward wording in some languages, but that it also plays into certain cultural biases while avoiding others. The level of difficulty and clarity of these tests may differ significantly for students from
different nations and speaking different languages.

Furthermore, there are issues with the structure of test administration and sampling. Students take PISA at 15-years-old, which means different things based on the country a student is from (Bracey, 2009); is it fair to compare a 15-year-old from a country where formal education begins at four years old with a 15-year-old from a country where this age is seven? Darling-Hammond (2014) recognized the limitations in participant selection as well, saying that “there are sampling issues with each administration that could influence the overall results” (p. 2). The way these tests are structured to include certain ages or grade levels does not ensure a representative sample of each country.

Result interpretation. How we analyze TIMSS and PISA results is another source of dispute. Many fault the tendency of media to consider solely test averages; for example, Nixon and Barth (2014) assert that the current method of reporting TIMSS scores – considering average scores alone – “can result in invalid inferences about the relative quality of educational systems”, which can then “lead to negative consequences for teachers and students” (p. 65). The authors proposed that, to make international comparisons, we need to develop a better, more comprehensive way to analyze test results. Bracey (2009) also criticized comparing countries based off of their average scores, as this neglects the wide disparity among scores within each country. For example, while students from the U.S. have a lower average score than several OECD countries on PISA science, in 2006 we had “more than twice as many [high scorers] as any other OECD nation” (p. 2), seemingly speaking to the strengths that do exist somewhere within U.S. science education. Nevertheless, we also had the most low
scorers of any nation besides Mexico. Making a meaningful analysis of these test scores is clearly more complex than comparing raw averages.

Also making the interpretation of TIMSS and PISA science results a difficult task is the variability of participating nations. Koretz (2009) explained, “these averages are not representative of a clear comparison group” (p. 39). While many nations have consistently participated throughout the history of TIMSS and PISA, certain countries test sporadically, and new countries are continually added to the testing pool. One solution to this issue could be creating groups of nations so that U.S. performance is compared “with the performance of other countries that provide an informative contrast” (Koretz, 2009, p. 40). For example, Koretz (2009) suggested Japan and Singapore as reasonable to include alongside the U.S. because of their consistent participation and consistent high achievement. Australia and Canada might also be valuable comparators because of the cultural similarities between countries. Nixon and Barth (2014) also recognized this problem; what we interpret as a decrease in science performance may be simply because new countries that excel in the sciences have joined the exam pool. In such a case, students from the U.S. might actually be improving in science, but because of these new players, we are given the impression that this is not the case. These issues aside, even those who do consider PISA and TIMSS averages cannot agree, with some believing that they indicate the U.S. is doing okay, pointing out that scores have remained relatively constant over time, while others interpret them as indicative of a colossal educational deficit on the U.S., drawing focus to the fact that we are being outperformed by more and more nations every round (Cavanagh & Manzo, 2009).

**Criticism from supporters.** Even those who do utilize PISA and TIMSS results
have noted the limitations. Beese and Liang (2010) concluded their research with a note of caution, saying that we should question the validity of the content used in these exams and the design of both test questions and the exam as a whole. Because of long-held doubts surrounding these areas in the education community, they urged us to view PISA results with reserve, and suggested that we don’t make any sweeping reform to education or policy strictly based off of the numbers PISA gives us. Koretz (2009) asserted that, while explanations for disparities in science test scores can be found, “the explanations remain speculative”; these differences may be because of “intentional differences in content, differences in sampling of students, or unintentional factors that have not yet been identified” (p. 41). Cavanagh and Manzo (2009) also mentioned that we must see how other countries approach education not solely in an educational setting, but in all aspects of a child’s life. Adapting to the educational practices of a high-performing nation does not account for differences outside of the classroom. Beese and Liang (2010) criticized the limits in understanding and unpacking these results along similar lines, yet they did not hold a positive outlook towards finding a solution: “The differences in educational systems reflect social and economic differences that cannot be fully accounted for” (p. 274). It is clear, then, that the criticism of TIMSS and PISA results are extensive and rooted in complex matters.

Discussion

Differences in Effects Internationally

While the goal of this paper was to understand and elucidate ways to improve U.S. students’ science achievement through comparison to other countries, I have seen that this may not be a realistic task. Principally, few variables considered for their effects
on science achievement had a clear-cut relationship with science performance; I did not see that students in high-performing countries uniformly do something that the U.S. does not and, consequentially, uniformly outperform U.S. students on science. Instead, the effects of many of these variables were different within each nation. Kaya and Rice (2010), Beese and Liang (2010), and Milford et al. (2010), for example, showed that familial SES has a different effect on science performance depending on the nation considered. House (2012), Kaya and Rice (2010), Papanastasiou and Zembylas (2004), and Ainley and Ainley (2011) showed that student attitudes toward science influence their science performance in drastically different ways depending on the country they come from. House (2012) showed that how U.S. students use technology influences their science performance differently than how students from Japan use technology. Beese and Liang (2010) and Milford et al. (2010) showed that school type and the quality of science resources might slightly influence science achievement in the U.S., while this was not true for all countries. Beese and Liang (2010) and House (2012) also revealed that the type of science instruction does not have a straightforward relationship with science performance; in the U.S., students who spend more time in labs and with hands-on experiments perform worse on these exams, while students from other countries who spend more time on these activities sometimes perform better. Kaya and Rice (2010), Milford et al. (2010), and Beese and Liang (2010) showed that teacher experience influences students from different countries differently.

**Non-significant Variables**

I also saw that some variables researchers proposed to be relevant to science education were not so. For example, Drummond and Sauer (2014) showed that the time
students spend playing video games does not influence their science performance. While it had been hypothesized that the prevalence of video games in Western culture might play a role in U.S. students’ underperformance in science, this was not the case. Baker et al. (2004) and Kaya and Rice (2010) showed that, in general, time spent in science class was not correlated with increased test scores, though this had previously been considered a potentially significant factor. Thus, while these findings reveal ways in which we do not need to improve to better science education, they do not help us understand how we could improve.

Science Curricula

The studies considered here revealed no quantifiable connections between science curricula and science performance. Nevertheless, it is evident from the widespread criticism of science curricula in the U.S. that this is a relevant topic. One common critique I saw is that the U.S. curricula are focused too much on memorization. Osborne (2013) showed, however, that internationally, students who indicated on the post-test questionnaire that they spent more time on memorization performed better on PISA. These results may challenge the use of this test data for practical purposes. If students must spend more time on memorization than on practical science to perform better on these exams, to “improve” their science achievement, is that really what we should be striving for as an educational goal? Perhaps U.S. performance on PISA and TIMSS are not measures we should be using to work towards better science curricula.

However, science curriculum is not universally held accountable for U.S. science achievement. Nixon and Barth (2014) contested that a principal reason TIMSS rankings are not indicative of education quality is differences in curricula. It may not necessarily
be that one country has a “better” curriculum than the other. Instead, they mentioned
different curricular goals as one factor responsible for differences in test performance;
how can we state that U.S. students are doing “worse” on such an international science
test when each country may be purposefully educating their students in areas that the test
does not cover? Furthermore, TIMSS does not account for a differential content
organization; two curricula may essentially cover the same topics but in a starkly
different order. In such a case, fourth and eighth graders from one country may have
been exposed to certain science topics before their peers from a different country. These
are merely differences in timing, not necessarily a reflection on educational quality. It
may be fruitful to explicitly establish what U.S. science curriculum should ideally cover
before condemning it.

TIMSS versus PISA

Another source of dispute is U.S. performance on TIMSS considered against U.S.
performance on PISA. For example, both fourth graders and eighth graders did, on
average, better than the international average for the 2007 TIMSS science exam. In
2006, however, fifteen-year-olds from the U.S. performed significantly below the
international average in PISA science (Cavanagh & Manzo, 2009). While this is an
intriguing observation, I found relatively little information explaining the sudden shift the
U.S. has had in its international ranking.

On this matter, Koretz (2009) concluded that, when looking at science scores,
“one can only speculate about the reasons for the different view provided by TIMSS and
PISA” (p. 47). However, he asserted that we should not take the results of one test in
isolation and that while each test has its flaws, the variations between these tests should
be used for further study rather than be pitted against each other (Koretz, 2009). As seen previously, Schmidt and McKnight (1998) attributed this difference to curricula, yet their discussion was speculative in nature.

Cavanagh and Manzo (2009) also offered insight, attributing this inconsistency largely to differences between the tests. Aside from the difference in age at which students take these exams, TIMSS is a curriculum-based exam, aiming to more narrowly test students on their factual science knowledge. PISA, on the other hand, tests how students can apply their knowledge to real-world situations. Where U.S. elementary and middle school students are regarding science curriculum may not, and perhaps should not, be directly related to where high school students are regarding science application. Policy makers also point out that fifteen-year-olds in different countries can be in a variety of grades; while this represents freshman or sophomore year in the U.S., students from certain countries have, by this point, been separated into technical or specialized schools. Furthermore, most countries that participate in TIMSS are of significantly different SES than the U.S., while most countries participating in PISA are similarly industrialized and wealthy (Cavanagh & Manzo, 2009).

Nevertheless, there is a downward trend over time when considering TIMSS alone, effectively controlling for these variables. Consistently, fourth graders from the U.S. have performed better, ranking-wise, than eighth graders on science. Following the third round of TIMSS, the U.S. continued its trend of fourth graders performing near the top of their peer groups in science and eighth graders sitting at scores just above the international average. Only two countries outperformed U.S. fourth graders in science, while 15 received scores significantly lower than the U.S. did. By the eighth grade, 25
countries tested higher than us, and only four countries scored significantly lower (Schmidt & McKnight, 1998). Thus, there does appear to be a strong suggestion of decline of science performance over the span of early childhood and secondary education. It may be pertinent for policy-makers to recognize the ways U.S. early childhood classrooms are succeeding in science, and to consider revisions for education after this point. As we have seen, however, the participant pools for fourth grade TIMSS, eighth grade TIMSS, and PISA are not consistent; one must decide, then, whether or not these trends are worth considering.

**Implications for Policy**

There are certain conclusions reached by this research that do suggest potential targets for policy. Kaya and Rice (2010) and House (2012) revealed that countries in which more students were in science tutoring, or involved in extra-curricular science activities, outperformed the U.S. Thus, this may suggest that increasing U.S. students’ involvement in such experiences could benefit their science achievement. I say “may suggest” because, as we’ve seen here, trends that hold between a handful of nations may not hold everywhere. These studies both looked exclusively at the U.S. compared to East Asian countries, so we do not have information on how tutoring or extra-curricular science affects science achievement elsewhere in the world. In particular, we do not know how it affects science achievement within the U.S. A more relevant study may be to compare students within the U.S. that are involved in these activities versus those who are not, and to see how this affects science achievement.

In fact, the conclusions reached by Kaya and Rice (2010) echoed these sentiments: in seeking to reform science education, policy makers should not merely
emulate those aspects of other countries outperforming the U.S. in the sciences. Rather, they should notice the disparities within the U.S. and seek to repair influencers of science achievement. Milford et al. (2010) reached similar conclusions, asserting that “factors appearing to be relevant in one nation cannot be assumed to work in another” (p. 470).

As this comparison has revealed, cross-cultural comparisons are no simple task. Perhaps a more effective level of change can be attained by playing to the U.S.’s strengths, and by modeling those schools within our country that are achieving high in science, as opposed to emulating other countries that are performing well in science. For as we’ve seen, factors that are correlated with one nation performing well on science may actually be correlated with another country performing badly in science; were we to note that other countries who spend more time with experimental-based instruction do better on TIMMS, and use this observation to implement more time spent on such instructional techniques, this may actually negatively impact U.S. students’ achievement in science. On the other hand, perhaps the issue here lies not with the time spent in such activities, but rather the nature and effectiveness of the activities themselves. The differentiation between quality and quantity of science instruction is clearly a complicated matter with more variables at play than any study, considered at this point, has comprehensively identified.

While most variables impacted different countries’ science performance differently, there were relatively uniform, international correlations with both TIMSS and PISA science scores and national wealth and familial structure. However, the implications this has for improving science education in the U.S. may be limited; policy makers cannot determine the family a child is born into. While the marital status of a student’s parents is not a likely target for improving science education, disparities in SES
U.S. SCIENCE PERFORMANCE INTERNATIONALLY

may be. However, this has seemingly shifted the focus of this issue from the science classroom to social justice; how we remedy this disconnect is certainly an interesting discussion.

**Implications for Future Research**

One potentially interesting area of future research is the connection between science achievement and student attitudes. As shown by Papanastasiou and Zembylas (2004) and House (2012), high-achieving science students in the U.S. commonly have negative outlooks on the subject. While developing ways to encourage positive outlooks toward science may thus not significantly improve U.S. TIMMS or PISA scores, this is still highly relevant, as students’ interest in and attitudes toward science have long-term implications beyond performance on PISA or TIMSS. The Organization for Economic Cooperation and Development has stated that science achievement, as measured by a test score, should not be the sole goal when developing science education (Ainley & Ainley, 2011). Rather, educators should foster student interest in science, as interest engenders effectual attitudes towards participation in real-world science issues, influences ultimate career paths, and encourages lifelong learning (Ainley & Ainley, 2011). How will high-achieving science students from the U.S. be encouraged to enter critical STEM professions with negative feelings toward science?

Another point of future research is the use of technology. Drummond and Sauer (2014) and House (2012) provided potentially conflicting conclusions on the effects of technology on science achievement. While House (2012) concluded that increased usage of computers for internet browsing and extracurricular purposes were correlated with lower science scores, Drummond and Sauer (2014) found there to be no relationship
between time spent on video games and PISA science scores. I cautiously label these results “potentially conflicting” because while they suggest opposite things about the quantity of technology use and science achievement, they are certainly not parallel studies. For one, computer usage and internet browsing, as defined in the first study, do not differentiate between gaming and other computer activities. Thus, it is possible that as students spend more time with technology, on top of time spent gaming, they may begin to perform worse in science. Furthermore, it is possible that these results reveal something about the nature of these tests. For example, as TIMSS is designed to be more factual based, increased time spent on computers for extracurricular purposes may detract from study time, thus resulting in lower test performance. On the other hand, students who spend more time gaming and with technology may develop problem solving skills or spatial reasoning that help them perform well on the “real-world” application design of PISA. Also significant here is that in both the U.S. and Japan – a top-performing science nation – increased use of computers for science learning was associated with lower scores. It may be beneficial to reevaluate how we use technology to teach science. This topic is especially relevant because of the widespread nature of technology in today’s society.

House (2012) showed that, internationally, students spending more time on memorization performed better on these exams. While this could, then, be an easy way to improve PISA or TIMSS scores – devoting more time to memorization – this may not be a desirable change to make; as discussed previously, the U.S. science curriculum is already seriously berated for the amount of memorization expected of U.S. students. This also begs the question of how valid are either of these claims. If students who perform
well on TIMSS and PISA indicate that they spend significant time on memorization, and researchers critique U.S. curricula for the overwhelming amount of memorization students do, then shouldn’t these students be performing better than they are? A relevant study could be done to quantify the time spent on these activities, rather than merely looking at the responses from post-test questionnaires.

Throughout my research, I came across an immense variety of possible factors contributing to differential science performance. I did not find any studies investigating one area that I hypothesize may contribute to this: U.S. students’ understanding and mastery of other subjects. Many studies scrutinized science classrooms, from how much time was spent in the class to what resources were available, and many studies looked at an expansive spread of factors outside of schools, from student variables to variables at the national level. Nevertheless, I contend that a pertinent area of study could be what else U.S. schools and teachers are doing, aside from science. It may be that U.S. science classrooms are comparable, on some level, to science classrooms in top-performing countries like Finland, Canada, or Japan. Perhaps the difference lies, then, in how else U.S. students are educated. For example, how are U.S. youth doing in mathematics education? It may be that a solid foundation in this subject is lacking, and without a strong foundation in math, U.S. students lack the tools that their international peers have when tackling science problems. Alternatively, other countries might better prepare their students in the arts, equipping them with better reading skills, with alternative research and writing methods, with a separate set of tools that could be vital for approaching problem-solving. In TIMMS and PISA both, word-problems are plethoric. Perhaps, then, U.S. science performance cannot be solely attributed to ill-preparation in the
sciences, but rather, it may be that students possess lower level reading and reasoning skills, thus making the actual test-taking process more difficult for them. I think this serves as an interesting, yet potentially difficult, point of further research.

**Final Thoughts**

The goal of this literature review was to investigate potential reasons that other countries outperform the U.S. in science; unearthing such reasons could serve as an exciting and useful tool for enacting policy to improve science education. Nevertheless, throughout the process of this research, it became increasingly clear that international science performance, as measured by TIMSS and PISA, is a complex matter. Simply emulating classroom or curricular practices followed by high-performing nations may not be an effective way to improve science education in the U.S. As practices that influence high science achievement in one country often do not have the same effects in another, it is clear that there are inextricable factors at play that ultimately impact student learning. Furthermore, those factors which do appear to have a relatively consistent relationship with science performance internationally (e.g. national wealth, familial structure) may not be feasible policy targets.

It is also clear that there is widespread debate about two relevant factors: the validity of TIMSS and PISA results, and science curricula in the U.S. Because of this, it may be more beneficial for schools and policy makers to reach a consensus about the goals for U.S. science curricula, and to evaluate these goals in line with international tests like TIMMS and PISA. Defining science learning outcomes in this way could help reveal if TIMSS and PISA are measures we should continue to use to evaluate student science performance in the future.
References


