New Waves Educational Research & Development
Vol. 22, No. 1, August, 2019, pp. 27–54

Gender Differences in General Achievement in Mathematics: An International Study

Ehsan Ghasemi
Hansel Burley
Parviz Safadel
Texas Tech University

Abstract

Women's underrepresentation in mathematics-related careers continues to concern policymakers, economists, and educators. This study addressed the issue by examining data from two international databases, namely IEA's *Trends in International Mathematics and Science Study 2015*, and the World Economic Forum's *Global Gender Gap Report 2017*. Using country as the unit for our observations and meta-analysis techniques, the question of gender mathematics differences was investigated using standardized mean difference comparisons and variance ratios. Fourth- and eighth-grade girls and boys were also compared in terms of the number of students who reached the advanced international benchmark. The findings mostly supported previous findings in the related literature; no statistically significant large differences were observed comparing the performance of girls and boys in mathematics achievement and the number of high achievers. Moreover, boys were found to have more variability in mathematics achievement than girls. This finding further isolates the potential cause of women's underrepresentation in mathematics-related careers as policies driven by implicit and explicit cultural biases.

Introduction

Ongoing concerns persist about the reasons women are underrepresented in mathematics-related careers despite efforts to keep the male-dominated job doors open for women (Frome, Alfeld, Eccles, & Barber, 2006; Wang, Eccles, and Kenny, 2013). The United States (U.S.) Department of Commerce, Economics and Statistics Administration's 2017 update on the status of women in science, technology, engineering, and mathematics (STEM) careers revealed that although women filled 47 percent of all U.S. jobs, they only held 24 percent of STEM jobs in 2015. Despite both an increase in STEM degrees granted and increases in STEM hiring, the U.S. continues to have a shortage of STEM workers. Finding ways to recruit more women into STEM careers is perceived as a viable solution (Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007).

Previous studies of mathematics performance among school age children show a decrease in the difference between boys' and girls' mathematics performance overall (Halpern et al., 2007; Hedges & Nowell, 1995; Hyde, Fennema, & Lamon, 1990; Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Lindberg, Hyde, Petersen, & Linn, 2010). The changes in education policies like the No Child Left Behind law and accountability movement might have contributed to the smaller gender gaps in mathematics (Cimpian, Lubienski, Timmer, Makowski, & Miller, 2016). The other speculation for the diminishing gender difference in mathematics performance, that is of interest to the authors, includes the closing of gender gaps due to cultural shifts and social changes (Else-Quest, Hyde, & Linn, 2010; Feingold, 1992). The ever-changing nature of socio-cultural,

economic, and political gender parities creates a need for more up-to-date cross-national analyses of gender differences in mathematics achievement, and mediators of these differences. In addition, using large, valid, and respected international databases that are representative of the population could result in more generalizable findings. In the related literature, few research studies assessed gender performance based on country classifications in terms of gender gap. Including the depth of economic and sociocultural gender disparity in a cross-national study of gender performance can provide additional understanding of factors associated with gender performance differences (Reilly, 2012; Reilly, Neumann, and Andrews, 2017).

Rationale and Goals of the Study

The present study was conducted to pursue several objectives including the investigation of gender differences in the average mathematics achievement, exploring gender differences in the number of high achievers in mathematics tests, testing the male variability hypothesis, and analyzing the effects of socio-cultural gaps on the direction and magnitude of gender differences in mathematics achievement.

One goal of the present study was to examine academic performance differences between girls and boys. The comparison included both the mean achievement comparison as well as the comparison of the number of high achievers. The measure of high achievement was advanced international benchmark (AIB) in the Trends in International Mathematics and Science Study (TIMSS). AIB is the highest benchmark in TIMSS mathematics examinations and is equal to the score of 625. Unlike previous studies that used percentiles to identify high-achieving students we adopted AIB as our measure. Hedges and Nowell (1995), for instance, collected data between 1960 to 1992 and found more males than females in the top 5% and 10% of the distribution. These researchers found a ratio range of 1.33 to 2.34 (Hedges and Nowell, 1995). The utilization of percentiles as a test statistic could be an appropriate strategy when a study includes multiple tests with varying assessment characteristics (*viz.*, different achievement benchmarks, test difficulty levels, time limits). However, when an identical test is administered to different samples, as in the case of TIMSS, using percentiles could be misleading. For instance, there might be some students at a high percentile in a country's sample who have only reached the low benchmark. For the same test in another nation, students reaching the highest benchmark might not be represented in the high percentiles because they are in the pool with more talented students. Therefore, unlike previous studies, we adopted the AIB as a common criterion for identifying high achieving and talented students.

The other goal of the study was to test the greater male variability hypothesis cross-nationally. In the literature on gender differences in mathematics achievement, some researchers have focused on variability hypothesis as one explanation for the higher number of males in STEM careers (e.g. Lindberg, et al., 2010; Machin & Pekkarinen, 2008; Penner, 2008; Hyde & Mertz, 2009). First introduced by Ellis (1974) in his study of institutionalized men and women with severe intellectual challenges, the greater male variability hypothesis posits that men represent more variability than women on several personality traits and psychological constructs (Benjamin, 1990; Hollingworth, 1914; Johnson, Carothers, & Deary, 2008; Thorndike, 1910). However, some previous research findings imply that higher male variability is mutable (e.g. Hyde & Mertz, 2009; Makel, Wai, Peairs, & Putallaz, 2016). Stevens and Haidt (2017) in a review of variability hypothesis literature posit that the gender difference in variability has reduced substantially over time within the United States. In addition, they assert that educational programs along with cultural and social factors can affect variability.

When applied to the gender difference in mathematics performance, the assumed higher male variance brings about a greater likelihood of male overrepresentation in upper and lower tails of a given construct distribution. In the present study, one of the goals was to compare the number of girls and boys in the upper end of the mathematics achievement distribution. The rationale for this comparison was to check the variability hypothesis; more male variability implies more boys reaching the highest scores. This approach focuses only the upper tail of the distribution and those who achieved the highest scores.

The other motivation for the current study came from scrutinizing TIMSS 2015 report and Global Gender Gap Report 2017 by the authors. The report on gender differences in mathematics performance across nations revealed that the held assumptions about gender disparity and mathematics performance (as assumed in gender stratification hypothesis) may not always hold true. In other words, based on our observation, we hypothesized that the direction and magnitude of gender differences in mathematics achievement might be the opposite of gender stratification hypothesis predictions when based on countries' rankings and their position in gender gap indices.

In sum, using TIMSS as the data source, country as the unit of our observation, and fourth and eighth grade mathematics performance as the explanatory variable, the current study was conducted to answer the following questions:

1. Is there a salient gender difference in average mathematics achievement among 4th and 8th grade students cross-nationally?
2. Is there a significant gender difference in the number of high achieving students in mathematics achievement?
3. Do boys have more variability than girls in mathematics achievement?
4. How does general gender parity ranking explain boys' and girls' mathematics achievement?

Theoretical Framework

Gender stratification hypothesis guided this study. This framework includes a variety of overlapping perspectives and research foci that share the causal assumption that a complex web of socio-environmental factors like power, privilege, and agency influence occupational outcomes (Fiorentine, 1993; Kane, 1992). One version of the hypothesis was advanced by Baker and Jones (1993) in that anticipated future opportunities shape current performance. Baker and Jones (1993) conducted a cross-national study of eighth graders to investigate the size and direction of gender differences in mathematics performance and how the levels of gender stratification of opportunity were correlated with gender differences in mathematics performance. They assumed that more opportunities for women would reshape the socialization, motivations, and social factors that decrease gender differences in mathematics performance, while they were in school. They used schooling and the labor market accessibility for women as two central institutional domains that increase opportunities for women, consequently creating more gender parities. They found significant variation in academic performance, that is, that boys did not outperform girls across all countries. This variation correlated with variation in access to higher education and job opportunities. The indicators with significant correlations were percentages of females in higher education, industrial work, labor force, and service jobs. Additionally, boy mathematics performance superiority for eighth graders in mathematics decreased from 1964 to 1986 in nine countries. In sum, Baker and Jones (1993) assert that "parity in opportunities for adults yields parity in preparatory performances." They found their findings as a challenge for biological and cognitive differences (e.g. Bock & Kolakowski, 1973; Levy, 1976; Stafford, 1961; Waber, 1979).

In this study, we have adopted and tested Baker and Jones (1993) perspective. We have also adopted the World Economic Forum's Global Gender Gap Report (2017) definition of gender gap; it defines gender gap as disparities across four thematic dimensions including economic participation and opportunity, educational attainment, health and survival, and political empowerment. In addition, it analyses the dynamics of gender gaps across industry talent pools and occupations. Baker and Jones (1993) claim that gender inequality and gaps could be a factor that negatively influences mathematics performance. They associated the variation in the gender stratification of educational and occupational opportunities in adulthood with cross-national variation in the mathematics performance across genders.

Literature Review

Various explanations regarding women's underrepresentation in STEM includes biological and evolutionary differences (Baron-Cohen, & Benenson, 2003; Geary, 2010; Levine, Huttenlocher, Taylor & Langrock, 1999) and gender stereotypes (Gunderson, Ramirez, Levine, & Beilock, 2012). Other reasons for the underrepresentation of women include girls' and teachers' mathematics anxiety (Beilock, Gunderson, Ramirez, & Levine, 2010; Maloney, Ramirez, Gunderson, Levine., & Beilock, 2015) and societal values that nudge girls into prioritizing the learning non-math related skills (Eccles, 1994; Hill, Corbett, & St Rose, 2010). Below, some of these trends are explained and their shortcomings are mentioned in brief.

One potential explanation for gender disparity in mathematics performance and achievement research has been biological and genetic differences. Penner (2008) puts biological explanations into three categories including genetic, hormonal, and cerebral explanations. He states that these explanations tend to focus on measurements of general intelligence and its components, especially spatial abilities. The study of gender difference in components of intelligence is also prolific. For instance, boys are assumed to be better in terms of spatial skills and abilities, while girls perform better in terms of verbal skills (Baron-Cohen & Benenson, 2003; Guiso, Monte, Sapienza, & Zingales, 2008; Halpern, et al., 2007; Hedges & Nowell, 1995; Wai, Lubinski, & Benbow, 2009. However, recent research findings are inconsistent. For instance, Spelke (2005) concluded that although mathematical and scientific reasoning abilities are developed from a set of biological cognitive capacities, boys and girls both have these capacities. Thus, boys and girls consequently develop equal talent for mathematics and science. Further, Linn and Hyde (1989) reviewed the literature and reported that gender differences in psychological and cognitive abilities is both small and declining. Therefore, the speculations based on evolutionary and biological differences appears to be insufficient in explaining the gender differences in academic performance and achievements.

A set of other studies have focused on factors such as attitudes, confidence, and values (Eccles & Wang, 2016; Ganley& Lubienski, 2016), thinking processes and strategy use (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998), teacher mathematics anxiety and stereotypes (Beilock, Gunderson, Ramirez, & Levine, 2010) and workplace culture and society reactions to women in male-dominated fields (Heilman & Okimoto, 2007; Heilman, Wallen, Fuchs & Tamkins, 2004). These factors have been found to contribute to gender difference in mathematics performance and mathematics-relevant career choice.

Fennema et al. (1998), for instance, conducted a case study of problem solving and computational strategies used by boys and girls as they progressed from the first grade to the third grade. Despite differences in strategies the boys and girls implemented, they found no significant gender difference in the number of correct solutions for a majority of the problem types. Their

findings revealed that girls tended to use more concrete strategies like counting and modeling while boys utilized more abstract strategies that reflected their conceptual understanding. Beilock et al. (2010) conducted a study to investigate the effect of female mathematics teachers' anxiety on the mathematics achievement of first and second grade students in the United States. They found that unlike the beginning of the school year when there was no relation between the two variables, at the end of the school year, more mathematics anxious teachers were more likely to have female students (not male students) who held the stereotype that "boys are good at math and girls are good at reading." The girls in the mathematics anxious teachers' classes who endorsed this stereotype also had lower mathematics achievements than girls who did not endorsed this stereotype and lower than boys overall. This set of studies targeted very young students with limited mathematics training and knowledge. Moreover, the occupational tendencies and decisions are normally made as students get closer to the end of compulsory education.

There are also some other researchers who have shifted the focus to workplace culture in male-dominated careers. Heilman et al. (2004), for instance, found that women success in male gender-typed jobs are less liked and more personally derogated compared to the equivalent success by men. They also found that being disliked can affect the overall evaluation and recommendations concerning organizational reward allocation. They argue that their findings provide support for the idea that gender stereotypes can cause bias in evaluative judgments of women even when these women have been successful.

Eccles and Wang (2016) by drawing of Expectancy-Value theory of achievement choices examined aptitudes and motivational beliefs as predictors of students move towards STEM occupations. They found that factors such as occupational and lifestyle values, math ability self-concepts, family demographics, and high school course-taking were stronger predictors of gender differences in the likelihood of entering STEM careers than math scores on an aptitude test. Eccles and Wang conducted their study on the students from one state in the U.S. They acknowledge that nations differ in factors such as when students start to make choices about which courses to take and other educational decisions. They claim that these factors affect students' career opportunities like decisions to specialize in STEM subjects versus other discipline areas.

Some other researchers have focused on socio-cultural environments to explain gender differences in scholastic achievements. These studies have utilized large international databases to alleviate the sample size and generalizability issues in other studies (Guiso et al., (2008); Else-Quest, et al., (2010). Relatively, the socio-cultural perspective has received more attention than biologically based explanations, recently. Guiso et al. (2008), for instance, assessed PISA (Program for International Student Assessment) 2003 data from 40 countries and found a positive correlation between gender inequality and gender gap in mathematics performance. They concluded that by increasing gender equality, it is possible to close the gender gap in mathematics. However, they acknowledge that there may be several other unobserved factors that affect the gender gap in mathematics; as Else-Quest, et al. (2010) claim "…it does not shed light on the specific domains of gender equity that are most relevant to mathematics achievement and leaves the debate about the mechanisms of the gender stratification hypothesis unresolved" (p 107). We find this perspective worthy of further investigation especially because of our observation regarding TIMSS 2015 results, the latest wave of TIMSS administration at the time of this study, which contradicted with findings of Else-Quest, et al. (2010).

In IEA's (International Association for the Evaluation of Educational Achievement) TIMSS 2015 report, of the 10 countries with the highest gender gap for fourth graders, we observed that girls who lived in eight countries performed better than boys in mathematics achievement tests

(Appendix 1). Four of these eight gender differences were statistically significant, including Saudi Arabia, Jordan, Bahrain, and Kuwait. Moreover, out of the 10 high-gap countries, as cited by Global Gender Gap Index (GGGI), nine countries had girls outperforming boys in the eighth grade. In two of these nine samples (i.e. Jordan and Bahrain), the differences were statistically significant. These observations imply that the economic, educational, and political disparities among genders positively correlate with girls' achievement in mathematics in several countries mostly located in the Middle East. As previously mentioned, one of the goals of this study is the investigation of the connections of socio-cultural, educational, economical, and political disparities in mathematics achievement cross-nationally with a lens on countries on the top and at the bottom of the gender gap scale. To achieve this, we conducted meta-analyses, both separately for each stratum of the distribution, as well as all countries together.

Method

Both TIMSS and Global Gender Gap Index are respected and reliable international databases. They are timely and provide a solid base for tracking changes and comparison internationally. The other advantage of these datasets is using data collection instruments consistently across different nations and different waves. The large and diverse samples that they used alleviate the issues of generalizability.

We utilized the effect size and meta-analysis techniques to makes gender comparisons. In case of large sample sizes, a small difference among two groups could result in statistically significant *p* value (Greenwald, Gonzalez, Harris, & Guthrie, 1996). As Kaplan, Chambers, and Glasgow (2014) point out many people prefer the treatment with a large than sample size while the effect size for the smaller trial must be larger to achieve the same significant *p* level. Therefore, the calculation of the effect size for large international databases seems warranted.

This section includes the description of the instruments and measures used for the study, the sample size and participating countries, and data analysis techniques and procedures.

Global Gender Gap Index

First introduced in 2006 by the World Economic Forum, the GGGI provides a picture of gender-related inequalities and tracks their changes over time by using a fixed methodology. The GGGI 2017 report includes 144 countries and their rankings on four dimensions including economic participation and opportunity, educational attainment, health and survival, and political empowerment (World Economic Forum, 2017). The rankings are based on women's disadvantaged status. Each of these four dimensions include several variables. For instance, the educational attainment variable includes the ratio of female literacy rate over male value, the ratio of female net primary enrollment rate over male value, the ratio of female net primary enrollment rate over male value, the ratio of female net secondary enrollment rate over male value, and the ratio of female gross tertiary enrollment over male value. World Economic Forum (2017) claims that the index has three features which makes it a good fit as a gender stratification gauge. First, GGGI measures gender-based gaps in relation to having access to resources and opportunities. Second, GGGI provides a picture of the outcomes as it is related to the indicators of the gender gap. Finally, the index focus is on the proximity of gender parity rather than women empowerment and determines if the gender gap has declined or increased. Due to its validity and previous use of the data as a measure of socio-economic indicator of gender gap, this report was utilized in the current study.

TIMSS 2015

TIMSS is a series of international assessments of science and mathematics knowledge. It was established by the IEA and are administered every five years since 1995 (IEA, 2017). One key objective of TIMSS is allowing educational systems worldwide to compare students' educational achievement and learn from their collective experiences with the goal of designing effective education policies. The TIMSS 2015 international database related to the mathematics tests was used in the current study. It is available to the public online.

Achievement benchmarks. The TIMSS database has four international benchmarks that provide information on participants' mathematics knowledge and skills measured on achievement scale points. The four points along the scale (i.e. benchmarks) are advanced international benchmark (the score of 625), high international benchmark (the score of 550), intermediate international benchmark (the score of 475), and low international benchmark (the score of 400). Based on IEA's descriptions, the achievement scale for mathematics ranges from 0 to 1000; however, most scores fall between 300 to 700. The mean of overall achievement was set to 500 and 100 was set as the standard deviation in 1995. See Appendix 2 for the percentages of fourth-grade students reaching each benchmark as reported in TIMSS 2015 report (IEA, 2017)

To create performance indices, TIMSS employs the scale anchoring analysis, a procedure that first identifyies the items the participant reaching international benchmarks answered correctly. Then, items are scrutinized to identify the reasoning skills and knowledge that answering the item correctly demonstrates. All the skills and knowledge for each benchmark are summarized in a list including competencies at each level. For instance, the list of students at and above international advanced benchmark includes 65 competencies. Some of the competencies include solving a multi-step reasoning problem involving division, solving a multi-step problem involving two-place decimals and whole numbers, and identifying a two-placed decimal on a number line marked with one-place decimals.

Sample size and participating countries. The general cross-national meta-analysis of fourth graders included 125,848 girls and 131,103 boys in forty-eight countries (i.e., k=48 effect sizes). The same meta-analysis for eighth graders included 127,351 girls, 126,813 boys, and forty countries. Following are the sample sizes when the original sample was divided into three categories of high gap countries, low gap countries, and in-between countries.

The meta-analysis for fourth graders from countries of high gender gap included 28,796 girls and 29,748 boys coming from nine countries. For fourth graders of low gender gap countries there were 26,077 girls, and 27,295 boys coming from ten countries.

The meta-analysis of eighth graders of high gender gap countries included 36,943 girls and 37,505 boys from ten countries. The meta-analysis of low gender gap countries included 17,388 girls and 16,868 boys coming from six countries.

For countries between high and low gap countries, there were 70,975 girls and 74,060 boys coming 39 countries including the United States. Regarding eighth graders in the same group there were 73,020 girls and 72,440 boys from 24 countries.

Data Analysis

The analysis is based on the meta-analysis procedures and techniques proposed by Lipsey and Wilson (2001). However, instead of different studies, nations were the units of analyses. Random

effect assumption was adopted as the starting point for the analysis, but both random and fixed effects were calculated and reported. The mathematics achievement data come from TIMSS (2015) while the gender parity measures come from Global Gender Gap Index (GGGI, 2017). Since TIMSS (2015) included fewer countries (i.e. 56 for all measurements) than GGGI 2017, the TIMSS' countries were used as the base for selecting high and low gap countries

In addition to including all countries and their corresponding achievement data for fourth and eighth graders, a set of countries were selected because of their high and low gender disparity as based on information available in TIMSS 2015 and GGGI 2017. The selected countries, with high gender parity as based on the GGGI 2017 report and the TIMSS 2015 report are as follows: Norway (2), Finland (3), Sweden (5), Slovenia (7), Ireland (8), New Zealand (9), France (11), Germany (12), Denmark (14), and Canada (16). The selected countries with the lowest gender parity are: Islamic Republic of Iran (140), Saudi Arabia (138), Morocco (136), Jordan (135), Turkey (131), Qatar (130), Kuwait (129), Bahrain (126), United Arab Emirates (120), and Republic of Korea (118). The numbers in the parenthesis stand for the countries' gender gap ranking with higher numbers representing more gender gaps. To check the possible difference in effect size variability of countries not included in these two groups, separate meta-analyses were conducted for the remaining countries for both fourth and eighth graders. In other words, we put our sample data into three groups including high gap countries, low gap countries, and countries in between, in addition to analyzing them altogether. Based on these, we hypothesized that the effect sizes of these three sets of countries were heterogeneous.

For the purpose of meta-analysis, several statistical packages and programs were utilized including Microsoft Excel, IBM SPSS (v24), IEA's IDB Analyzer, and SPSS macros by Wilson (2005). IDB Analyzer was used to generate SPSS syntax that converted multiple TIMSS data performance estimates to an aggregated plausible value. The syntax also produced standard errors that were adjusted for the complex sampling techniques used to select participants.

Using the aggregated plausible value, the procedures proposed by Lipsey and Wilson (2001) were followed to calculate the first statistics of interest, the effect sizes (*d*) and effect size standard errors.

The basic formula for d is $d = \frac{M_{girls} - M_{boys}}{SD_{pooled}}$, where *d* equals the performance mean of girls minus the performance mean of boys, divided by the pooled standard deviation.

In the analyses regarding mathematics achievement, positive values of *d* showed the superiority of girls, and negative values of *d* showed the superiority of boys. We used Cohen's (1988) rules of thumb for interpreting the absolute value of *d*, with 0.2 as a small effect size, 0.5 a medium effect size, and 0.8 a large effect size. Effect sizes below 0.2 were considered trivial. There are some arguments, however, about this cut off point with some scholars arguing that studies should be interpreted by their practical significance (e.g. Baguley, 2009).

Next, after aggregating the *d*s, an analog to the ANOVA was conducted to check whether each set of *d*s was consistent across countries and shared a common effect size using both fixed and random effects analysis. Tests for homogeneity analyses produced Q, τ^2, and I^2 statistics. Q was assessed using the χ^2 distribution with k-1 degrees of freedom, where k is the number of countries in the analysis. The τ^2 statistic, an adjustment of Q produced from random effects analysis, was evaluated based upon a rule of thumb that values greater than 1 indicated heterogeneity. The I^2, a ratio of true heterogeneity ($I^2 = \frac{Q-(k-1)}{Q} * 100$) was evaluated on a rule of thumb that values of 25%, 50%, and 75% indicate low, medium, and high variability, respectively (Higgins & Thompson, 2002; Higgins, Thompson, Deeks & Altman, 2003).

The second statistic used for analysis was the variance ratio, a statistic that can support understanding findings from effect sizes (Halpern et al., 2007). Procedures and techniques proposed by Feingold (1992) were followed in calculating and cumulating of variance ratios for each country. The within gender standard deviation for the girls and boys of each country was squared to determine the variance. Boys' variance was divided by the girls' variance to calculate the variance ratio for each country. The variance ratio close to 1 (i.e., 0.9 to 1.1) indicated gender variance parity, while values greater than 1 indicated more variance of among boys. Values less than 1 indicated more variance among girls. (Feingold, 1994; Reilly, Neumann, & Andrews, 2017).

To compare girls and boys in terms of reaching and passing the advanced international benchmark, we employed the third statistic of interest, odds ratios. The odd ratios were based on two by two contingency tables and are defined as the odds for reaching the benchmark for girls relative to odds for reaching the benchmark for boys. The countries with fewer than five participants in contingency table cells were removed from the analysis. Argentina, Kuwait, and Saudi Arabia were removed from the fourth-grade analysis, and Argentina, Morocco, and Saudi Arabia were removed from the eighth-grade analysis because of having values less than 5 in the contingency tables (Agresti, 2003). The logs of the odd ratios were used to convert odd ratios into *d*s (Borenstein, Hedges, Higgins, & Rothstein, 2011). The positive values of *d* showed higher number of girls while negative values of *d* indicated more boys in the samples.

Like previous studies (Hyde, et al., 2008; Guiso, et al. 2008) we also calculated the male to female ratios. We followed the procedure that Hedges and Nowell (1995) utilized in their analysis of mental test scores from six studies of national probability samples to compute the ratios of the estimated number of boys and girls who fell into the high benchmark category.

Findings

Tables 1 and 2 and Figures 1 and 2 represent descriptive statistics of the countries investigated in the study and the distribution of unweighted effect sizes around grand mean. The tables include number of female and male students, their means in mathematics achievements, their corresponding standard deviations, and unweighted effects sizes (i.e. Cohen's *d*s*)*. For fourth graders, 44 countries out of 48 countries (91.66%) had effect sizes below 0.2. The countries with effect sizes above 0.2 were Bahrain (0.207), Saudi Arabia (0.482), Oman (0.22), and Italy (-0.28). For eighth graders, the effect sizes of 37 countries out of 40 countries (92.5%) were below 0.2. The countries with effect sizes above 0.2 included Botswana (0.231), Oman (0.340), and Chile (-0.225). The United States' effect sizes for fourth and eighth graders were -0.09 and -0.02 respectively.

Table1. *Unweighted Effect Sizes and Descriptive Statistics of all Countries in the Study (Fourth Graders)*

Country	1(f)	N2(m)	Mean1(f)	Mean2(m)	SD1(f)	SD2(m)	*d*
Bahrain	2194	2042	459.02	441.52	79.21	89.52	0.207
Iran	1863	1960	429.03	418.27	96.74	104.6	0.107
Korea Repub.	2258	2411	604.18	611.63	64.89	69.44	-0.11
Kuwait	1782	1805	357.51	344.7	98.06	103.7	0.127
Morocco	2479	2589	380.33	378.68	89.81	92.93	0.018
Qatar	2547	2647	440.38	437.57	90.98	102.5	0.029
Saudi Araba	2181	2156	405.42	362.51	80.53	96.64	0.482
UAE	10314	10860	453.41	449.93	100.9	109.2	0.033

Turkey	3178	3278	482.15	484.12	91.61	98.65	-0.02
Canada	6035	6226	506.08	515.15	73.39	76.16	-0.12
Denmark	1833	1877	535.56	541.67	73.94	76.18	-0.08
Finland	2437	2578	540.13	530.79	64.18	68.54	0.141
France	2392	2480	485.03	491.14	73.14	75.36	-0.08
Germany	1895	2047	519.8	523.27	64.28	66.34	-0.05
Ireland	2060	2284	545.08	549.36	70.86	75.05	-0.06
New Zealand	3074	3248	489.39	491.7	85.65	93.33	-0.03
Norway	2146	2183	550.92	547.28	68.16	72.82	0.052
Slovenia	2151	2294	517.67	521.97	64.61	72.39	-0.06
Sweden	2054	2078	519.11	518.28	68.72	69.45	0.012
Australia	2937	3972	512.69	522.09	8150	85.17	0
Bulgaria	2076	2145	526.99	522.01	81.73	83.21	0.06
Chile	2310	2415	458.08	459.68	71.04	75.02	-0.02
Chinese Taipei	2088	2203	593.73	599.35	67.47	73.73	-0.08
Croatia	1937	2048	496.13	508.26	63.86	67.6	-0.18
Cyprus	2015	2057	520.96	528.55	76.4	81.68	-0.10
Czech Repub.	2578	2615	524.42	531.71	69.5	70.06	-0.10
Georgia	1896	2020	464.93	461.72	84.4	88.74	0.037
Hong Kong	1614	1979	609.16	619.05	64.01	66.68	-0.15
Hungary	2506	2530	526.28	532.03	85.63	90.1	-0.07
Indonesia	1928	2096	392.56	385.96	88.33	92.1	0.073
Italy	2115	2245	496.75	516.65	69.75	71.89	-0.28
Japan	2198	2182	593.07	592.62	65.73	71.6	0.007
Kazakhstan	2312	2390	545.66	543.24	81.57	82.99	0.029
Lithuania	2250	2273	536.56	534.12	68.29	74.05	0.034
Oman	4513	4568	436.63	414.66	97.53	102.5	0.22
Netherlands	2209	2225	525.84	533.93	55.06	57	-0.14
Poland	2360	2385	534.06	535.6	68.15	74.18	-0.02
Portugal	2293	2399	535.63	546.55	79.5	73.89	-0.14
Russian Fed.	2442	2476	564.19	563.68	72.53	72.86	0.007
Serbia	1967	2062	519.6	516.87	82.13	90.67	0.032
Singapore	3178	3335	619.52	615.94	84.06	87.78	0.042
Slovak Repub.	2802	2966	492.5	503.63	78.73	80	-0.14
Spain	3774	3954	499.19	510.97	66.64	71.3	-0.17
USA	5035	4908	535.76	542.95	79.66	83.21	-0.09
England	1973	1918	542.96	549.5	80.45	87.7	-0.08
North. Ireland	1509	1601	569.23	571.37	85.39	86.05	-0.02
Belgium	2713	2650	542.92	548.7	60.42	61.17	-0.1
Argentina	1447	1443	418.13	423.57	80.83	79.97	-0.07

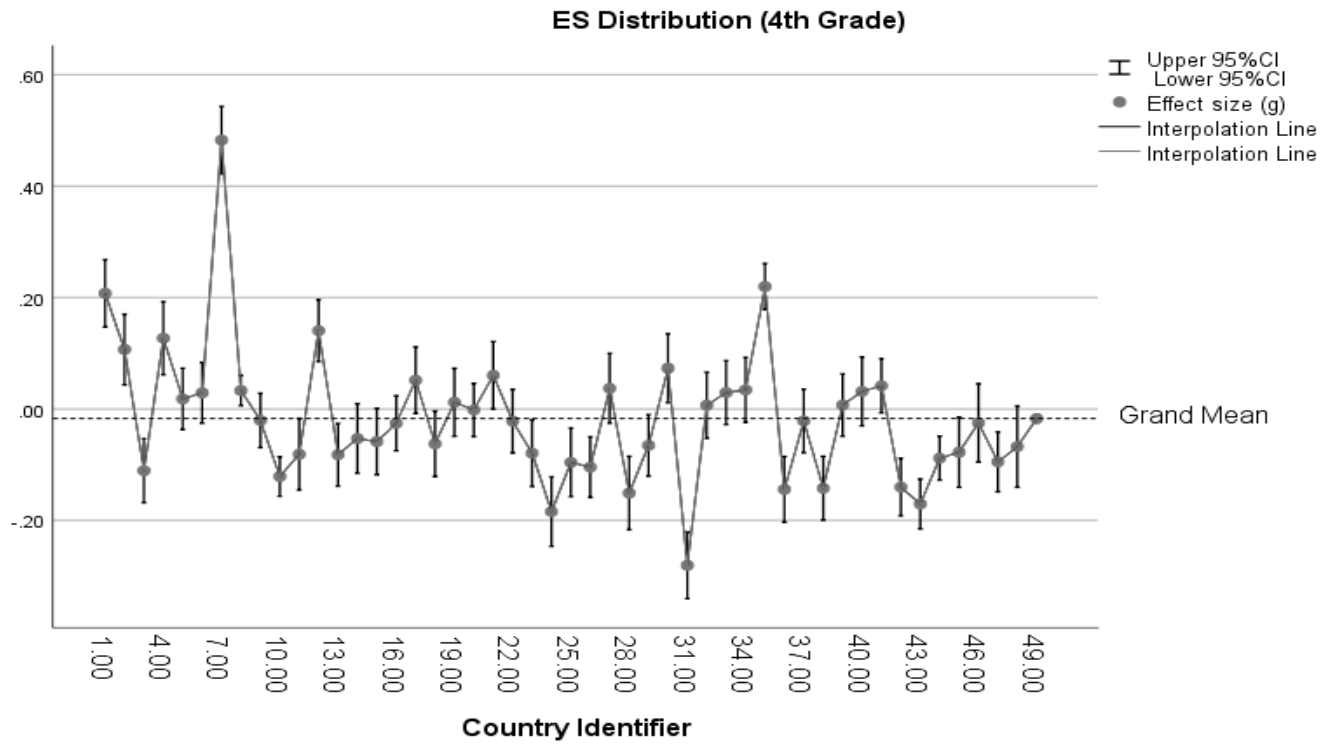


Figure 1. Distribution of Unweighted Effect Sizes Around the Grand Mean (Fourth Grade).

1. Bahrain; 2. Iran; 3. Korea Republic; 4. Kuwait; 5. Morocco; 6. Qatar; 7. Saudi Araba; 8. UAE; 9. Turkey; 10. Canada; 11. Denmark; 12. Finland; 13. France; 14. Germany; 15. Ireland; 16. New Zealand; 17. Norway; 18. Slovenia; 19. Sweden; 20. Australia; 21. Bulgaria; 22. Chile; 23. Chinese Taipei; 24. Croatia; 25. Cyprus; 26. Czech Republic; 27. Georgia; 28. Hong Kong SAR; 29. Hungary; 30. Indonesia; 31. Italy; 32. Japan; 33. Kazakhstan; 34. Lithuania; 35. Oman; 36. Netherlands; 37. Poland; 38. Portugal; 39. Russia; 40. Serbia; 41. Singapore; 42. Slovak Republic; 43. Spain; 44. USA; 45. England; 46. N. Ireland; 47. Belgium; 48. Argentina

Table 2. *Unweighted Effect Sizes and Descriptive Statistics of all Countries in the Study (Eighth Graders)*

Country	N(f)	N(m)	Mean(f)	Mean(m)	SD(f)	SD(m)	*d*
Bahrain	2296	2618	462	446.6	72.53	86.16	0.193
Iran	2999	3128	437.95	434.84	89.77	97.79	0.033
Jordan	4278	3585	394.85	376.26	89.07	97.37	0.199
Korea Republic	2605	2703	605.04	606.38	80.94	89.03	-0.016
Kuwait	2132	2044	395.97	388.71	82.51	98.15	0.080
Morocco	6194	6751	385.43	383.81	79.87	80.14	0.020
Qatar	2670	2719	440.4	434.07	96.25	107.86	0.062
Saudi Arabia	1998	1761	374.73	360.38	78.54	92.81	0.167
UAE	8838	9073	471.05	458.91	90.75	104.31	0.124
Turkey	2933	3123	461.14	454.73	104.71	105.93	0.061
Canada	4389	4229	525.83	530.76	67.02	71.53	-0.071
Ireland	2408	2264	520.84	526.31	71.5	76.3	-0.074
New Zealand	4239	3762	494.24	491.88	83.67	92.48	0.027

Norway	2340	2328	511.07	512.62	67.99	70.87	-0.022
Slovenia	2049	2199	515.09	517.58	68.92	69.54	-0.036
Sweden	1963	2086	497.8	504.49	70.9	72.26	-0.093
Australia	5125	4974	504.45	506.9	82.04	82.33	-0.030
Botswana	3026	2927	400.27	381.1	78.43	87.16	0.231
Chile	2305	2505	418.37	436.23	78.42	80.29	-0.225
Chinese Taipei	2794	2912	599.07	599.15	94.09	100.09	-0.001
Georgia	1931	2098	453.73	452.81	87.39	95.86	0.010
Hong Kong	1968	2178	591.52	596.77	72.9	83.03	-0.067
Hungary	2471	2419	509.74	519	92.49	94.07	-0.099
Israel	2691	2796	509.98	512.82	97.91	105.42	-0.028
Italy	2213	2255	490.92	497.63	73.14	75.84	-0.090
Japan	2417	2325	587.56	585.41	87.45	90.38	0.024
Kazakhstan	2387	2500	531.05	524.63	92.31	94	0.069
Lebanon	2075	1792	441.25	443.87	74.32	76.34	-0.035
Lithuania	2101	2238	509.94	512.64	76.12	78.51	-0.035
Malaysia	5014	4710	470.01	460.66	84.56	88.42	0.108
Malta	1878	1930	495.07	492.72	85.29	90.75	0.027
Oman	4365	4492	419.92	387.76	88.35	100.25	0.340
Russia	2294	2484	533.16	542.59	82.38	80.83	-0.116
Singapore	2977	3136	625.74	616.55	76.94	86.39	0.112
South Africa	6422	6084	375.85	368.78	86.91	87.1	0.081
Thailand	3509	2973	439.54	421.98	85.42	92.45	0.197
Egypt	4010	3803	396.6	387.38	97.76	99.18	0.094
USA	5091	5071	517.6	519.29	81.15	85.3	-0.020
England	2391	2351	521.14	517.35	81.29	77.43	0.048
Argentina	1565	1487	389.25	401.21	89.59	88.68	-0.134

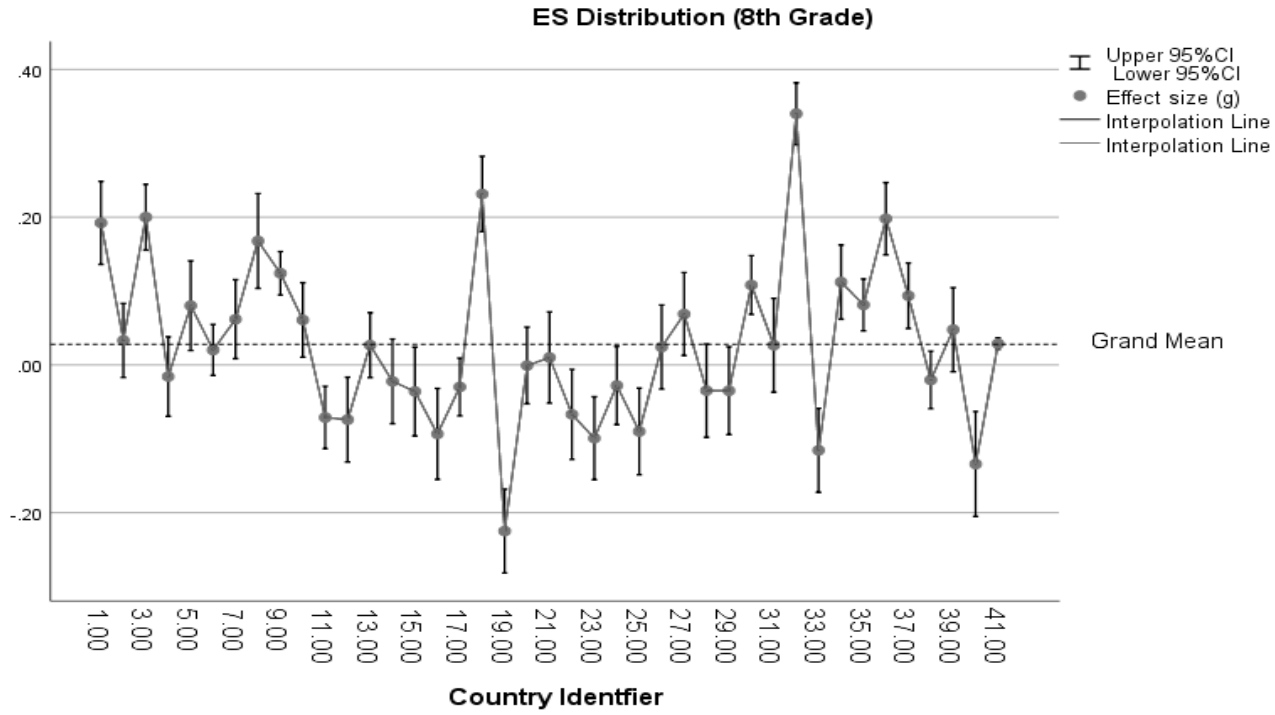


Figure 2. Distribution of the Gender Difference Unweighted Mathematics Achievement Effect Sizes Around Grand Mean (Eighth Grade).

1. Bahrain; 2. Iran; 3. Jordan; 4. Korea Republic; 5. Kuwait; 6. Morocco; 7. Qatar; 8. Saudi Arabia; 9. UAE; 10. Turkey; 11. Canada; 12. Ireland; 13. New Zealand; 14. Norway; 15. Slovenia; 16. Sweden; 17. Australia; 18. Botswana; 19. Chile; 20. Chinese Taipei; 21. Georgia; 22. Hong Kong; 23. Hungary; 24. Israel; 25. Italy; 26. Japan; 27. Kazakhstan; 28. Lebanon; 29. Lithuania; 30. Malaysia; 31. Malta; 32. Oman; 33. Russia; 34. Singapore; 35. South Africa; 36. Thailand; 37. Egypt; 38. USA; 39. England; 40. Argentina

Gender difference in the Average Mathematics Achievement (Research Question 1)

Table 3 shows the results of meta-analyses related to the research question 1 (i.e. is there a salient gender difference in average mathematics achievement among 4th and 8th grade students cross-nationally?). The answer to this question is "no." The average weighted effect size of all fourth graders was d = -0.017. The effect sizes were heterogeneous, Q (47) = 892.59, p < .001, I^2 = 94%. The random effects τ^2 was .013. The average weighted effect size of the all eighth graders together was d = 0.028. The effect size showed heterogeneity, Q (39) = 757.6, p <.001, I^2 = 95% The random effects τ^2 was .011.

Table 3. *Overall Effect Size of Gender Differences in Average Math Achievement for Fourth and Eighth Graders*

						Test of Null		95% CI		Homogeneity Test		
4th Grade		k	ES	SE	V	Z	*P*	Lower	Upper	Q	*df*(Q)	P
	Fixed	48	-.016	.004		-4.25	.000	-.024	-.009	892.59	47	.000
	Random	48	-.017	.017	.013	-0.99	.319	-.051	-.016			

		K	ES	SE	V	Z	P	Lower	Upper	Q	df(Q)	P
8th Grade	Fixed	40	.043	.004		11.04	.000	.036	.051	757.6	39	.000
	Random	40	.028	.017	.011	1.62	.10	-.005	.063			

The Representation at High-performance Levels (Research Question 2)

Specifically, regarding research question 2 (i.e. is there a significant gender difference in the number of high achieving students in mathematics achievement?) the average weighted effect size for fourth grade was -0.068 (d = -0.068, Q (44) = 213, p < .001, I^2 = 79%). Figure 3 represents the range and frequency of the effect sizes.

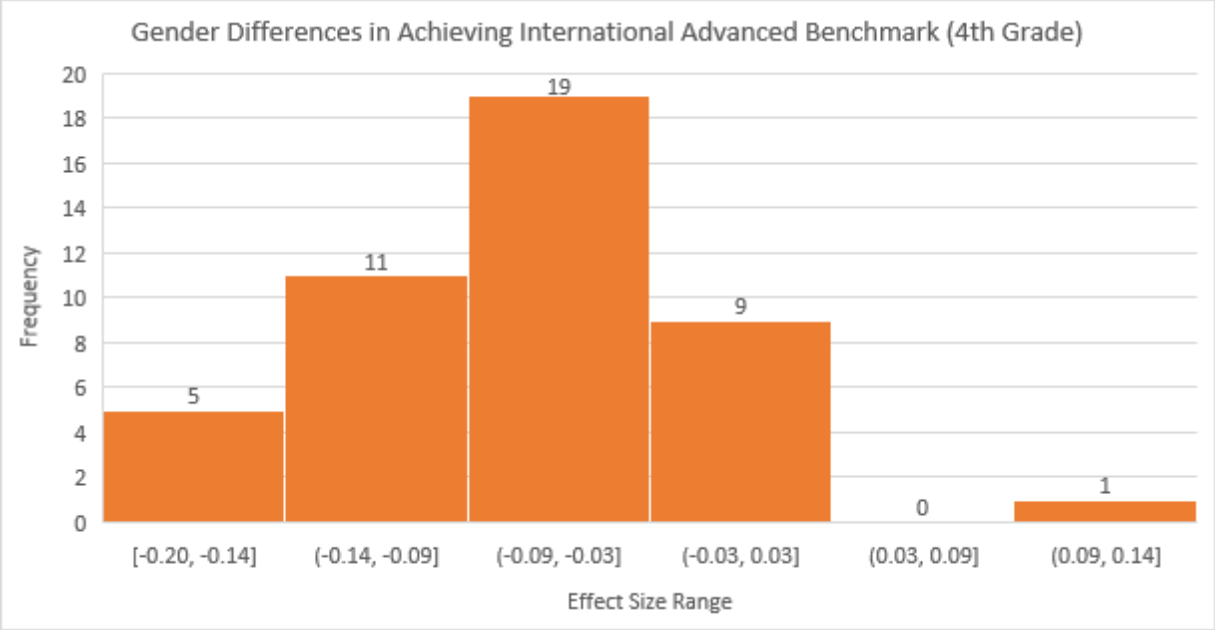


Figure 3. Distribution of Gender Difference Effect Sizes in Reaching the International Advanced Benchmark in Mathematics for Fourth Grade Students; Negative Values Indicate Higher Number of Boys.

The analysis of eighth grade data came to very similar findings. The average weighted effect size was -0.063 (Q (36) = 272.2, p < .001, I^2 = 86%). Again, the answer to question 2 is "no" based on the results. Figure 4 represents the range and frequency of effect sizes.

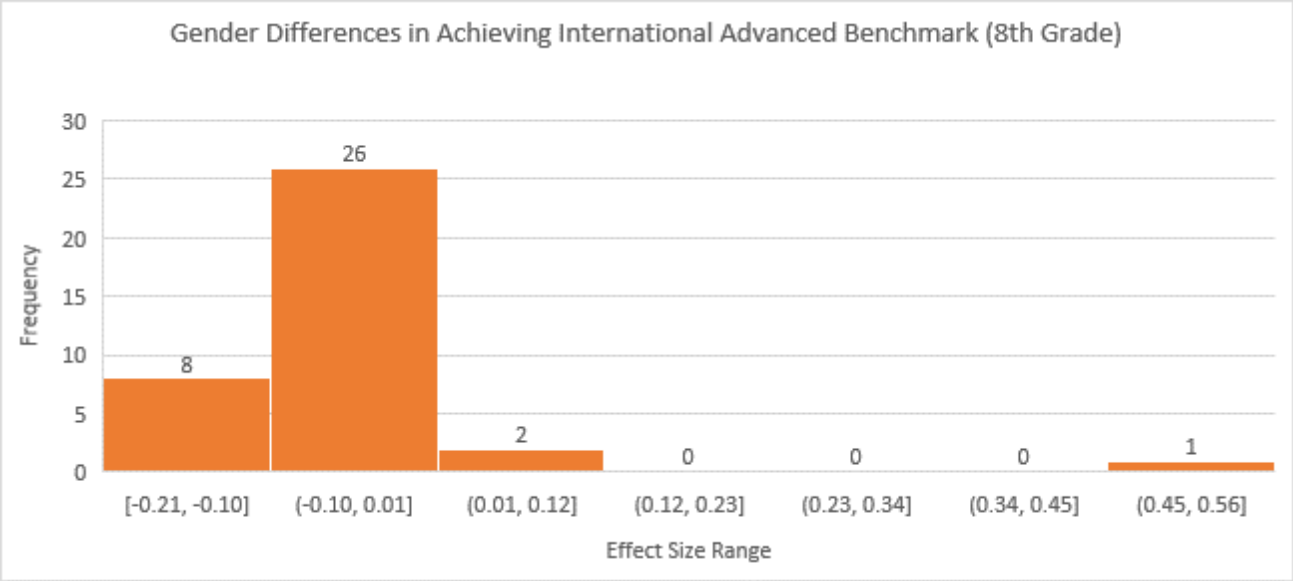


Figure 4. Distribution of Gender Difference Effect Sizes in Reaching International Advanced Benchmark for Eighth Grade Students; Negative Values Indicating the Higher Number of Boys.

The boy to girl ratios for the students who achieved AIB for the fourth grade ranged from 0.67 to 3.0. Of 49 countries assessed only five had values less than 1 (Argentina, Oman, Bulgaria, Finland, and Kazakhstan) indicating higher female ratios. The rest had values above 1. Eight ratios fell between 0.9 to1.1. (see Figure5)

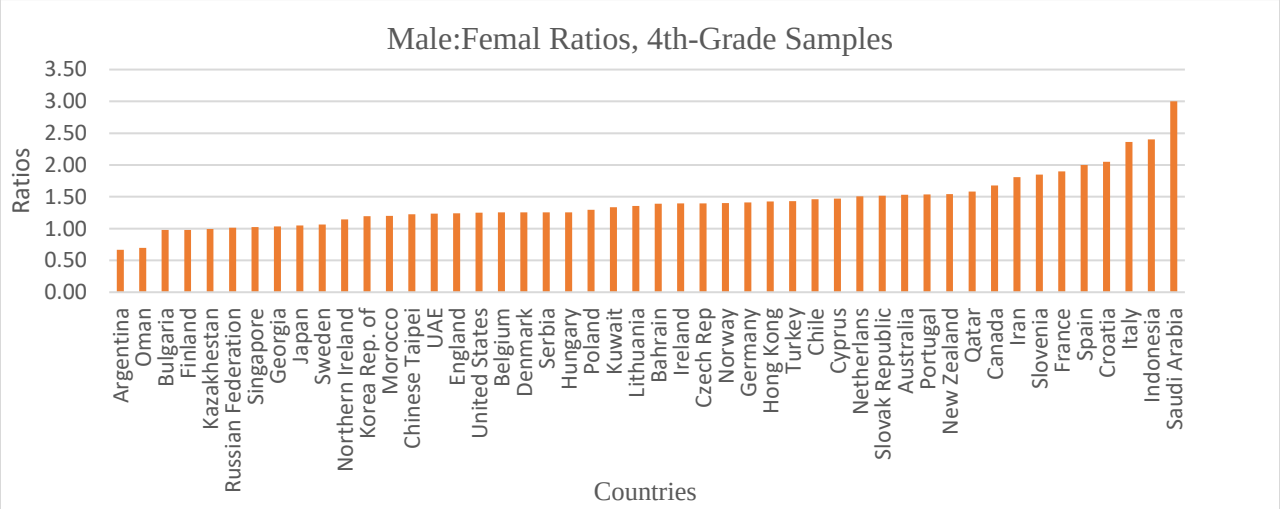


Figure 5. Male to Female Ratios of Fourth-grade Students Who Reached AIB.

The range of boy to girl ratios for eighth-grade samples was 0.90 to 4. Out of 40 countries, five had ratios less than 1 (Turkey, Singapore, Thailand, England, and Japan). Morocco had the ratio of one and the rest had ratios above one. Seven ratios were between 0.9 to 1.1 (Figure 6).

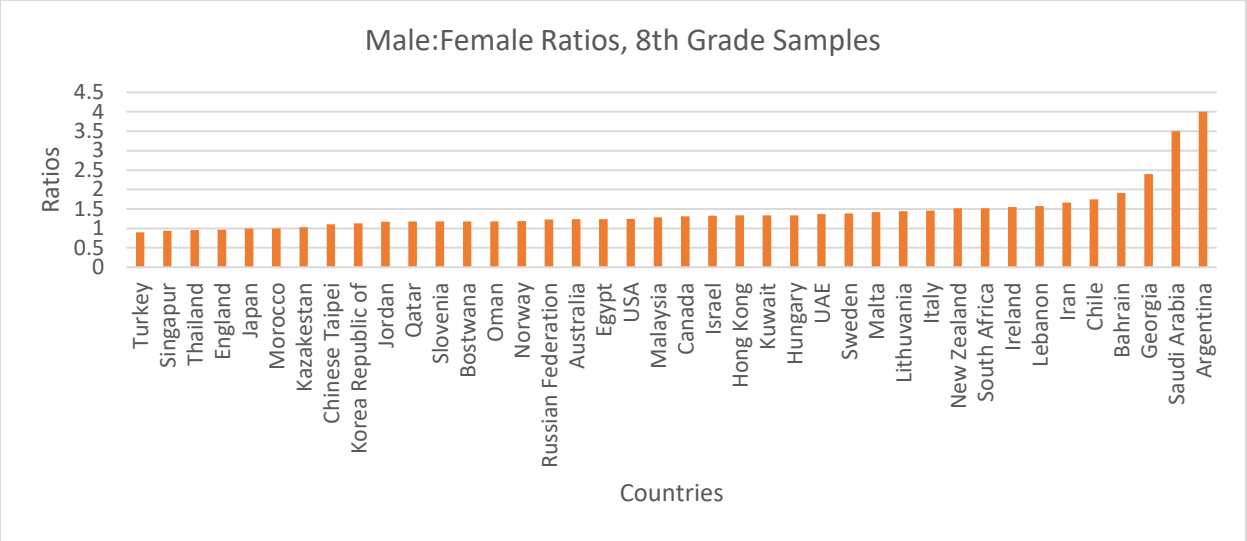
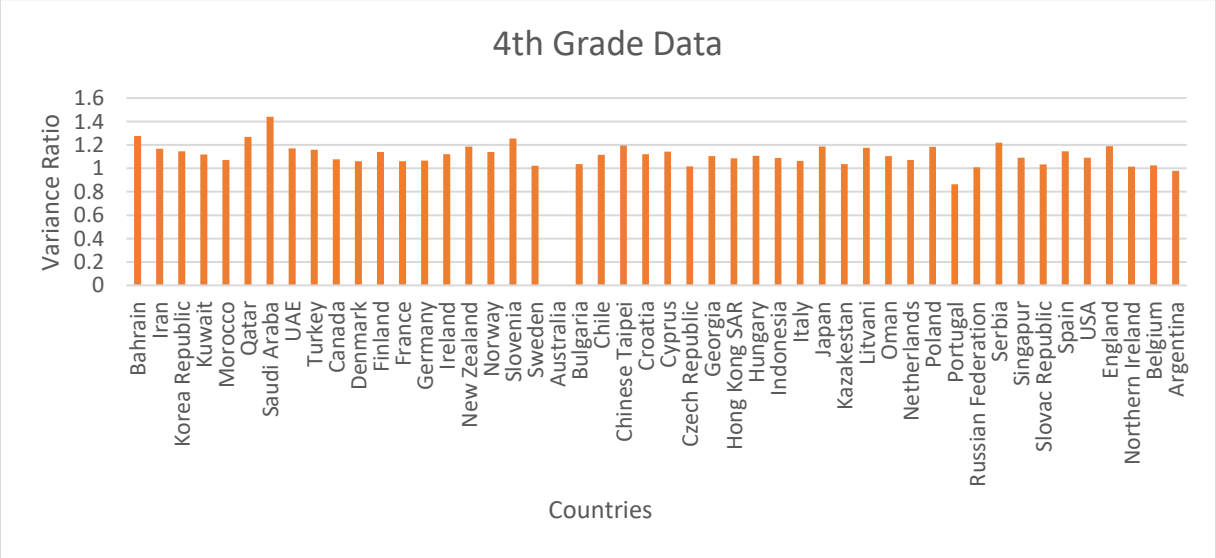


Figure 6. Male to Female Ratios of Eighth-grade Students Who Reached AIB.

Greater Male Variability Hypothesis (Research Question 3)

To answer the research question 3 (i.e. do boys have more variability than girls in mathematics achievement?) VRs (Variance Ratios) for individual countries and the median were calculated. The mean VRs for the fourth and eighth grade data were 1.06 and 1.13, respectively; and the medians were 1.1 and 1.08 respectively.

Figure 7 represents the individual VR values for the fourth and eighth grades, respectively. Out of 48 VRs of fourth grade data, 45 of them (93%) were above 1, and 26 of them (54%) were higher than 1.1. Only the VR of Portugal was below 0.9 (0.86). Out of 40 VRs (i.e. countries) examining eighth grade data, 37 VRs (92%) were higher than 1 while 21 VRs (52%) were greater than 1.1. No VR was below 0.9.



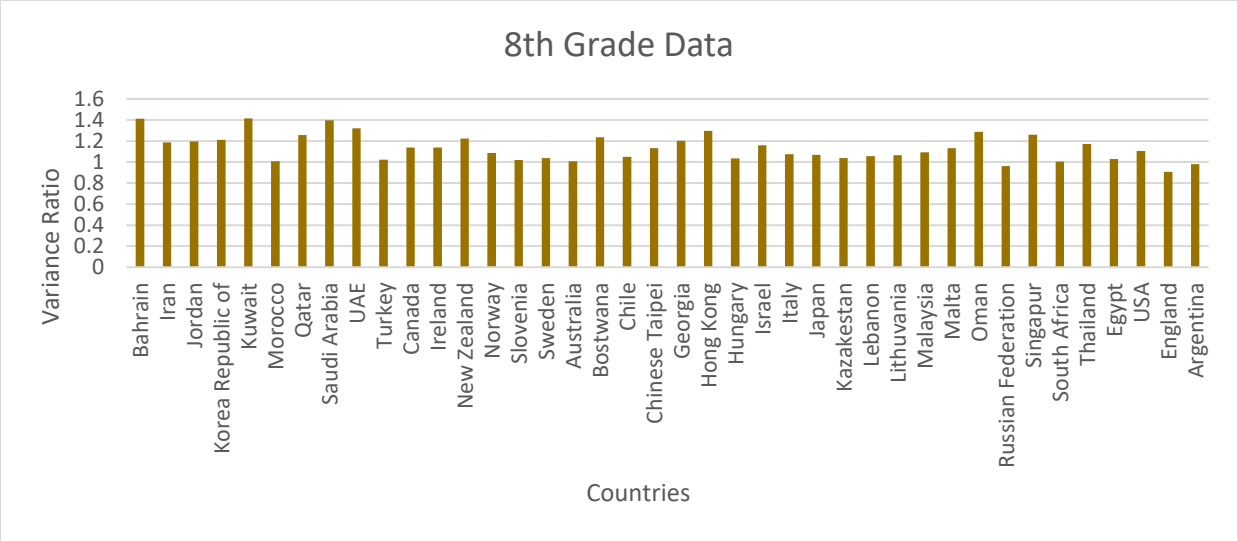


Figure 7. Variance Ratios of Boys to Girls in Mathematics Achievement (Fourth and Eighth Grades).

Gender Gap and Gender Difference in the Mathematics Achievement (Research Question 4)

Table 4 represents the results of meta-analyses related to the high-gap countries. The average weighted effect size (4th grade) was d = 0.096. The effect sizes were heterogeneous, Q (8) = 267.28, p < .001, I^2 = 97%. The random effects variance (τ^2) was 0.021. The average weighted effect size of eighth graders in high gap countries was d = 0.092. The associated Q-statistic indicated heterogeneity, Q (9) = 84.83, p <.001, I^2 = 89%. The random effects variance was .004.

Table 4. *Effect Size of Gender Differences in Math Achievement for Fourth and Eighth Graders in High-gap Countries*

						Test of Null		95% CI		Homogeneity Test		
4th Grade		k	ES	SE	V	Z	*P*	Lower	Upper	Q	*df*(Q)	P
	Fixed	9	.069	.008		8.40	.000	.053	.085	267.28	8	.000
	Random	9	.096	.050	.021	1.91	.054	-.002	.194			
8th Grade		K	ES	SE	V	Z	P	Lower	Upper	Q	df(Q)	P
	Fixed	10	.090	.007		12.38	.000	.076	.105	84.83	9	.000
	Random	10	.092	.023	.004	3.96	.000	.046	.137			

As represented in table 5, the results of meta-analyses of low-gap countries revealed that average weighted effect size of fourth graders in low-gap countries was d =- 0.028. The effect sizes were heterogeneous, Q (9) = 78.06, p < .001, I^2 = 88%. The random effect variance was 0.005. The average weighted effect size of eighth graders in low gap countries was d = -0.043. The effect size showed heterogeneity, Q (5) = 15.58, p =.008, I^2 = 67%. The random effects τ^2 was .001.

Finally, the average weighted effect size of fourth graders for the in-between countries was d =- 0.043. The effect sizes were heterogeneous, Q (28) = 406.17, p < .001, I^2 = 93%. The random

effects variance was 0.010. The average weighted effect size of eighth graders in this group was *d* = +0.039. The effect size showed heterogeneity, Q (23) = 554.95, *p* =.000, I^2 = 95%. The random effects τ^2 was .015.

Random effects categorical analysis (i.e. analog to the ANOVA, non-iterative method of moments) was conducted through SPSS Wilson's macros. The results showed that the six groups of effect sizes, including fourth graders in high gap, low gap, and in-between countries and eighth graders in low gap, high gap, and in-between countries were heterogeneous (Q = 25.98, *df = 5, p < .001,* I^2 = 80%*)*. This difference between the groups should be interpreted with cautions because of the low power of I^2 in the analyses with small number of effect sizes.

Table 5. *Effect Size of Gender Differences in Math Achievement for Fourth and Eighth Graders in Low-gap Countries*

						Test of Null		95% CI		Homogeneity Test		
4th Grade		k	ES	SE	V	Z	*P*	Lower	Upper	Q	*df*(Q)	P
	Fixed	10	-.039	.008		-4.57	.000	-.056	-.022	78.06	9	.000
	Random	10	-.028	.025	.005	-1.10	.269	-.079	.022			
8th Grade		K	ES	SE	V	Z	P	Lower	Upper	Q	df(Q)	P
	Fixed	6	-.040	.010		-3.72	.000	-.061	-.019	15.58	5	.008
	Random	6	-.043	.019	.001	-2.23	.025	-.081	-.005			

Discussion

The results suggest that despite some variabilities in the unweighted effect sizes of investigated countries, the difference between boys and girls in mathematic achievement is negligible when the homogeneity of variance is assumed and the mean difference is used as the statistic (Cohen, 1988). This finding is mainly consistent with studies of previous waves of TIMSS (Baker & Jones, 1993; Mullis, Martin, Fierros, Goldberg, & Stemler, 2000; Reilly, Neumann, and Andrews, 2017). For instance, in Else-Quest et al. (2010) analysis on TIMSS 2003 data, the overall effect size (*d*) for eighth graders in 46 countries was -0.01 which compared to the current study finding (*d*=0.02). While the unweighted effect sizes in Else-Quest, et al.'s study ranged from -0.42 (Bahrain) to 0.40 (Tunisia), the effect sizes of the current study ranged from -0.22 (Chile) to 0.34 (Oman) implying variation within a smaller range since the test administration in 2003. Eighth graders' overall effect size (fixed = 0.043, random = 0.028) showed both an increase and inclination towards girls' better performance compared to fourth graders (fixed = -0.016, random = -0.017). However, both are too small to conclude that gender differences in math performance increases as students grow and go to the high school (Hyde, Fennema, & Lamon, 1990).

Regarding students of the U.S. participating in this study, the effect sizes for the fourth grade and eighth grade were -0.09 and -0.02 respectively. It suggests a very small and negligible boys' better performance with no significant gender difference in math achievement. This finding compares to the results of TIMSS 2003 analysis (Else-Quest, et. al., 2010) in which *d* value for eighth grade students of the U.S. was 0.06 (positive value of *d*, in contrary to the current study, represented boys' better performance), TIMSS 2011 analysis (Reilly, Neumann, and Andrews, 2017) *d* =0.05, TIMSS 1995 analysis (Mullis et al., 2000) d = 0.05, and SIMS 1982 (Baker & Jones, 1993) d = -0.01. The findings are also comparable to the Hyde et al. (2008) study of state

assessment data on ten states; *d* values were -0.01 and -0.02 for fourth and eighth graders respectively (positive values indicated boys' better performance).

Among the studied countries, there were a few countries that stood out in terms of the magnitude of gender difference in the average mathematics achievement. For instance, girls from Saudi Arabia and Oman had higher achievement than boys with the effect sizes that were relatively larger than most of the studied countries. While these values could be considered as outliers, the other implication is the role of cultural characteristics and not psychological differences in gender differences in mathematics achievement. A support for this implication is the cultural similarities of Oman and Saudi Arabia and the point that both are Middle Eastern nations. In other words, gender differences in mathematics may be culturally and/or geographically shaped.

The VR values of the current study compare to findings of previous meta-analysis studies of both international databases (e.g. Reilly and Neumann, 2017, VR = 1.16) and previous gender studies (e.g. Lindberg, et.al, 2010, VR = 1.08) as well. Reilly and Neuman (2017) conducted their analysis by using TIMSS 2011 eighth grade student data and interpreted the VR of 1.16 as consistent with greater male variability hypothesis. Therefore, the current finding, while still consistent with greater male variability hypothesis (i.e. VR = 1.08), implies an eight percent decrease in male variability in mathematics achievement over the course of four years. Moreover, the current findings indicate more male variability in fourth grade (10%) than eighth grade (8%). These interpretations are made based on median values (see Shaffer, 1992 for the discussion of cautions in the interpretation of the VR). Regarding individual countries VR, the findings could be interpreted as further evidence of greater variability of boys in mathematics achievement in numerous countries.

While the current study provides evidence for greater male variability, one question is what magnitude of variance difference may create a significant difference in ratio of the number of boys to girls in the right tail of the distribution. Hyde (2014) by referring to Hedges and Friedman (1993), claim that variance ratios that are very close to 1, like the finding of the current study, cannot explain underrepresentation of women in STEM fields, especially because most of the STEM careers do not need mathematics skills at very extreme percentiles.

When countries were divided based on their gender gap, their gender difference mean effect sizes were still negligible (below 0.2). For the low gap countries, the mean effect sizes showed a decrease which indicated boys' had better achievements in mathematics compared to the overall population of all countries. The Q test value was significant, and I^2 value, despite a decrease, was still high (Borenstein et al., 2011). It was the opposite of general expectation in gender stratification framework.

The mean effect sizes of the high gap countries increased in favor of girls when compared to the overall effect size, cumulating data from all countries. These findings indicate at least two points: first, there is an increase in the mean effect sizes in favor of girls. Based on gender stratification hypothesis, it was expected to come up with mean effect sizes smaller than the overall effect sizes which favored male students in countries with less educational, economic, and political gender equalities and opportunities. And second, no significant change in the heterogeneity was observed as the values of I^2 (the proportion of true between-country variance to the total variance) were still comparable and the Q test values were still highly significant.

Therefore, regarding gender stratification hypothesis, the current study findings are not supportive. By dividing countries into categories based on socioeconomic standards, the meta-analysis revealed that the Global Gender Gap Index did not account for the between-country variability of effect sizes. This finding suggests that higher gender equity (political, economic,

etc.) is not necessarily positively correlated with higher parity in math performance for all countries. Reilly (2012) and Reilly et al. (2017) found similar patterns in PISA and TIMSS when they coded countries as either developing (non-OECD) and developed (OECD) though their values were statistically negligible. In addition, in the current study the mean effect size for counties of high standards of gender parity (in economy, education, etc.) were all negative values showing better performance of boys. The same statistic value for countries of low gender parity standards were all positive, implying girls' better math achievement. Their difference from the overall effect sizes were also the opposite of expected trend. However, the observed heterogeneity of effect sizes does not allow drawing the conclusion that the same effect is being measured across nations for each group. Moreover, for the small number of effect sizes, the I^2 is not a strong measure of heterogeneity (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella 2006).

One possible explanation for mixed findings in the related literature regarding gender stratification hypothesis is the nature of the data used to examine the hypothesis. There are instances of support for this hypothesis when PISA is used but not when TIMSS data is utilized. The Else-Quest et al. (2010) study on TIMSS 2003, Reilly, Neumann, and Andrews's (2017) study on TIMSS 2011, and the current study on TIMSS 2015 could not find support for the hypothesis in mathematics achievement. Reilly, Neumann, and Andrews (2017) make this point by attributing the contradictory findings from these two tests to the differences in the age of participants and the composition of participating countries in PISA and TIMSS.

The other implication of the current study is a need for more granularity in variables that are included in the analyses. This study has investigated gender differences in mathematics cross-nationally. As previously mentioned, one of the findings was high heterogeneity across nations. Although measures of socio-cultural and economic disparities were included in the analyses, this did not assist in reducing the heterogeneity of results. There is an intricate network of complex phenomena in each society that might not have been fully captured by the used instruments and study designs. As Hyde (2014) points out, scientists should beware of making global statements regarding gender differences referring to the entire nation or universal differences, and intersectionality can alleviate the issue; intersectionality is an approach that considers multiple categories of identity, difference, and disadvantage such as gender, race, class, sexual orientation, disability, and religion (Cole, 2009, as cited in Hyde, 2014).

The other possible consideration for further research is this field is the importance of difference in difficulty level of test questions. One of the speculations regarding women underrepresentation in STEM careers is that boys are better in solving mathematics problems that require complex skills. This speculation has not been supported in the United States (Hyde et al., 2008). However, future gender studies could focus on questions that require higher levels of skills in the international mathematics tests.

Talented Students and the Variability Differences

The variability hypothesis (also called greater male variability hypothesis) states boys have higher variability than girls in psychological traits. The hypothesis implication regarding cognitive abilities is that boys are more likely to have very high and very low intelligence due to the higher variability. In the twentieth century, the association between the male higher variability and the occurrence of higher number of boys with higher scores was invoked by Thorndike (1910) in his study of gender differences in variability in the cognitive tasks. However, Hedges and Nowell (1995) argue that the observed difference could have been due to greater boys' mean score, greater boys' score variance, or both. The focus of this study was on one aspect of cognitive ability, namely

the mathematics knowledge and skills. While the average mean difference was negligible, and the average variance was slightly higher for boys, the mean effect size of the number of male and female high achievers were not significantly different.

Among samples of two analyses (fourth and eighth grades), only four *d*s were 0.2 or higher which indicated statistically significant difference in the number of girls and boys who achieved the AIB (Oman = 0.52, Saudi Arabia = -0.27, Georgia = -0.21, Italy = -0.2). The number of countries with significant VRs were higher, however (more than half of the samples). In two samples together (i.e. fourth and eighth grades), there were 48 VRs above 1.1 and 17 VRs above 1.2. These numbers imply that the higher male variability and higher number of significant VRs, did not lead to higher number of boys reaching the benchmark. Considering the observed absence of gender difference in the central tendency, the source of the high variance of these individual countries could have come from the lower end of the distribution. This observation highlights the gap of a quantitative procedure to integrate *d* and VR to make an objective evaluation (Feingold, 1992).

Limitations of the Study

One of the main limitations of the study is our logic for deciding how many countries would go into the three categories: high gender parity, low gender parity, and in-between countries. The idea of putting 10 countries in high and low category came from World Economic Forum's tradition of announcing 10 countries with highest and lowest gender parity measures. The other issue arose when there were less countries in TIMSS than the Global Gender Gap report. Therefore, we were sometimes hesitant in selecting the closest countries to our criteria for which the data were available in both datasets. Moreover, the idea of separate analysis for each set of countries might have increased the possibility of type I error. The same analysis could be done with meta-regression techniques with gender gap variable as an independent variable. We implemented separate analysis mainly because we were interested in comparing the magnitude of the mean effect sizes.

The other limitation of the study is using Global Gender Gap Report indices to explain the between-country variations in the effect sizes. The random effect variance was small when all countries were considered in the meta-analysis. The dispersion of effect sizes as Figures 1 and 2 represent are mostly between 0.2 to -0.2 for both grades. One could criticize our further analysis based on gender gap indices and find them as unnecessary since there was no outstanding heterogeneity in the first place. Knowing this, we kept our post-analysis in the paper to provide a model check when gender gap was included as a factor.

Conclusion

All in all, the results of current meta-analysis study suggest that gender difference in mathematics in overall international population is negligible in terms of central tendency, and there is slightly greater variance in boys' scores in mathematics tests. However, there is no difference in the number of girls and boys who achieved the highest benchmark in the TIMSS 2015. Observing both negative and positive values of effect sizes, reveals the malleable nature of math achievement for girls and boys at least in elementary and middle school. Further studies might replicate the current study with some international database of students in higher levels of education (e.g. colleges and universities) to check how their performance compare. We believe our findings suggests that it is the society that stereotypes and causes the gender disparity. Our findings indicate that gender differences are either absent or very small and negligible during elementary and middle

school years. It is possible that as students grow up and get closer to making decisions on their education and career path, the differences become more significant and noticeable. Moreover, some more qualitative studies with adolescents and adults are warranted to explain what happens to girls' motivation to pursue STEM careers; the reasons that are not revealed through the numerical analysis.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported?.*British journal of psychology*, *100*(3), 603–617.

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education*, 91–103.

Baron-Cohen, S., & Benenson, J. F. (2003). Books and arts-essential difference: Men, women and the extreme male brain/essential difference: The truth about the male and female brain. *Nature*, *424*(6945), 132–132.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences*, *107*(5), 1860–1863.

Benjamin Jr, L. T. (1990). Leta Stetter Hollingworth: Psychologist, educator, feminist. *Roeper Review*, *12*(3), 145–151.

Bock, R. D., & Kolakowski, D. (1973). Further evidence of sex-linked major-gene influence on human spatial visualizing ability. *American journal of human genetics*, *25*(1), 1.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons.

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open*, *2*(4), 2332858416673617.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L. Lawrence Earlbaum Associates.

Cole, E. R. (2009). Intersectionality and research in psychology. *American psychologist*, *64*(3), 170.

Eccles, J. S. (1994). Understanding women's educational and occupational choices. *Psychology of Women Quarterly*, *18*(4), 585–609.

Eccles, J. S., & Wang, M. T. (2016). What motivates females and males to pursue careers in mathematics and science?. *International Journal of Behavioral Development*, *40*(2), 100–106.

Ellis, H. (1974). *Man and woman: a study of human secondary sexual characters*: New York: Arno Press.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological bulletin*, *136*(1), 103.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, *62*(1), 61–84.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex roles*, *30*(1-2), 81–92.

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational researcher*, *27*(5), 6-11.

Fiorentine, R. (1993). Theories of Gender Stratification: Assumptions, Evidence, and "Agency" and "Equity" Implications. *Rationality and Society*, *5*(3), 341–366.

Frome, P. M., Alfeld, C. J., Eccles, J. S., & Barber, B. L. (2006). Why don't they want a male-dominated job? An investigation of young women who changed their occupational aspirations. *Educational Research and Evaluation*, *12*(4), 359–372.

Ganley, C. M., & Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. *Learning and Individual Differences*, *47*, 182–193.

Geary, D. C. (2010). Male, female: the evolution of human sex differences (2nd ed.). Washington, DC: American Psychological Association.

Greenwald, A., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated?. *Psychophysiology*, *33*(2), 175–183.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, *320*(5880), 1164–1165.

Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex roles*, *66*(3-4), 153–166.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51.

Handelsman, J., Cantor, N., Carnes, M., Denton, D., Fine, E., Grosz, B., ... & Sheridan, J. (2005). More women in science. *Science*, *309*(5738), 1190–1191.

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research*, *63*(1), 94–105.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*(5220), 41–45.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: the implied communality deficit. *The Journal Of Applied Psychology*, *92*(1), 81–92.

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: reactions to women who succeed at male gender-typed tasks. *The Journal Of Applied Psychology*, *89*(3), 416–427.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, *21*(11), 1539–1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), 557.

Hill, C., Corbett, C., & St Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics*. American Association of University Women. 1111 Sixteenth Street NW, Washington, DC 20036.

Hollingworth, L. S. (1914). Variability as related to sex differences in achievement: A critique. *American Journal of Sociology*, *19*(4), 510–530.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index?. *Psychological methods*, *11*(2), 193.

Hyde, J. S. (2014). Gender similarities and differences. *Annual review of psychology*, *65*, 373–398.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly*, *14*(3), 299–324.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*(5888), 494–495.

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*(22), 8801–8807.

International Association for the Evaluation of Educational Achievement (IEA). (2015). TIMSS 2015 and TIMSS Advanced 2015 international results. Retrieved from: http://timss2015.org/

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, *3*(6), 518–531.

Kane, E. W. (1992). Race, gender, and attitudes toward gender stratification. *Social Psychology Quarterly*, 311–320.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: a cautionary note on the potential for bias. *Clinical and translational science*, *7*(4), 342–346.

Levine, S. C., Huttenlocher, J., Taylor, A., & Langrock, A. (1999). Early sex differences in spatial skill. *Developmental psychology*, *35*(4), 940.

Levy, J. (1976). Cerebral lateralization and spatial ability. *Behavior Genetics* (6)171–188.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological bulletin*, *136*(6), 1123.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational researcher*, *18*(8), 17–27.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, Calif.: Sage Publications.

Machin, S., & Pekkarinen, T. (2008). Assessment. Global sex differences in test score variability. *Science (New York, N.Y.)*, *322*(5906), 1331.

Makel, M. C., Wai, J., Peairs, K., & Putallaz, M. (2016). Sex differences in the right tail of cognitive abilities: An update and cross cultural extension. *Intelligence*, *59*, 8–15.

Maloney, E. A., Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2015). Intergenerational effects of parents' math anxiety on children's math achievement and anxiety. *Psychological Science*, *26*(9), 1480–1488.

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

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology*, *114*(S1), S138–S170.

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PloS one*, *7*(7), e39904.

Reilly, D., Neumann, D. L., & Andrews, G. (2017). Investigating Gender Differences in Mathematics and Science: Results from the 2011 Trends in Mathematics and Science Survey. *Research in Science Education*, 1–26.

Stevens, S., & Haidt, J. (2017). The greater male variability hypothesis-An addendum to our post on google memo. Retrieved from: https://heterodoxacademy.org/the-greater-male-variability-hypothesis/

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist*, *60*(9), 950.

Stafford, R. E. (1961). Sex differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual and motor skills*, *13*(3), 428–428.Terman, L. M. (1947). *Genetic Studies of Genius: 25 Years' Follow-up of a Superior Group. The Gifted Child Grows Up*. Stanford University Press.

The United States Economics and Statistics Administration, Department of Commerce. (2017). Women in STEM: 2017 Update. Retrieved from: https://www. commerce.gov/file/women-stem-2017-update

Thorndike, E. L. (1910). *Educational Psychology*. New York: Columbia University, Teachers College.

Waber, D. P. (1979). "Cognitive Abilities and Sex-related Variables in the Maturation of Cerebral Cortical Functions." Pp. 58–67 in *Sex-related Differences in Cognitive Functioning*, edited by M. A. Wittig and A. C. Petersen. New York: Academic Press.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817.

Wang, M. T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological science*, *24*(5), 770–775.

Wilson, D.B. Meta-analysis macros for SAS, SPSS, and Stata. (2005). Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

World Economic Forum. (2017). Global Gender Gap Report 2017. Retrieved from: https://www.weforum.org/reports/the-global-gender-gap-report-2017

Appendices

Appendix 1

Descriptive statistics of the fourth graders' mathematics achievement for ten countries with the highest gender gap adopted from GGGI 2017, and TIMSS 2015 reports

Country	Avg. Girls	Avg. Boys	Girls Higher	Boys Higher	Sig.
Saudi Arabia	405	363	-43		●
Jordan	384	368	-15		●
Bahrain	359	347	-12		●
Kuwait	437	426	-10		●
Iran	403	393	-10		
UAE	465	461	-3		
Qatar	537	534	-2		
Morocco	519	518	-1		
Turkey	569	571		2	
Korea, Republic of	526	534		7	●

Descriptive statistics of the eighth graders' from ten selected countries with the highest gender gap as based on the 2017 GGGI and 2015 TIMSS reports

Country	Avg. Girls	Avg. Boys	Girls Higher	Boys Higher	Sig.
Jordan	395	376	-19		●
Bahrain	462	446	-16		●
Saudi Arabia	375	360	-14		
UAE	471	459	-12		
Kuwait	396	389	-7		
Qatar	440	434	-7		
Turkey	461	455	-6		
Iran	438	435	-3		
Morocco	385	384	-2		
Korea, Republic of	605	606		1	

Appendix 2

Percentages of fourth-grade students reaching International Benchmarks, adopted from TIMSS 2015

Country	Advanced Benchmark	High Benchmark	Intermediate Benchmark	Low Benchmark
	625	550	475	400
Australia	9	36	70	91
Bahrain	2	13	41	72
Belgium	10	47	88	99
Bulgaria	10	40	75	92
Canada	6	31	69	92
Chile	1	10	42	78
Chinese Taipei	35	76	95	100
Croatia	3	24	67	93
Cyprus	10	39	74	93
Czech Republic	8	38	78	96
Denmark	12	46	80	96
England	17	49	80	96
Finland	8	43	82	97
France	2	21	58	87
Georgia	2	15	47	78
Germany	5	34	77	96
Hong Kong SAR	45	84	98	100
Hungary	13	44	75	92
Indonesia	0	3	20	50
Iran, Islamic Republic of	1	11	36	65
Ireland	14	51	84	97
Italy	4	28	69	93
Japan	32	74	95	99
Kazakhstan	16	47	80	96
Korea, Republic of	41	81	97	100
Kuwait	0	3	12	33
Lithuania	10	44	81	96
Morocco	0	3	17	41
Netherlands	4	37	83	99
New Zealand	6	26	59	84
Northern Ireland	27	61	86	97
Norway (5)	14	50	86	98
Oman	2	11	32	60
Poland	10	44	80	96
Portugal	12	46	82	97
Qatar	3	13	36	65
Russian Federation	20	59	89	98
Saudi Arabia	0	3	16	43
Serbia	10	37	72	91
Singapore	50	80	93	99
Slovak Republic	4	26	65	88
Slovenia	6	34	75	95
South Africa (5)	1	5	17	39
Spain	3	27	67	93
Sweden	5	34	75	95
Turkey	5	25	57	81
UAE	5	18	42	68
United States	14	47	79	95