

Sex Differences in Mathematics and Science Achievement: A Meta-Analysis of NAEP Assessments

David Reilly¹, David L. Neumann^{1,2}, Glenda Andrews^{1,2}
Griffith University

Note : This is a pre-print version of the manuscript for self-archiving. This article may not exactly replicate the final version published in the APA journal. It is not the copy of record, and pagination may be different.

Citation :

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, 107(3), 645-662. doi: 10.1037/edu0000012

Abstract

Gender gaps in the development of mathematical and scientific literacy have important implications for the general public's understanding of scientific issues and for the underrepresentation of women in science, technology, engineering and math (STEM)-related fields. Data from National Assessment of Educational Progress (NAEP) were subjected to a meta-analysis to examine whether there were sex differences in mathematics and science achievement for students in the USA across the period 1990-2011. Results show that there were small but stable mean sex differences favoring males in mathematics and science across the past two decades, with an effect size of $d = .10$ and $.13$ respectively for students in twelfth grade. Furthermore, there were large sex differences in high-achievers, with males being overrepresented by a factor of over 2:1 at the upper-right of the ability distribution for both mathematics and science. Further efforts are called for to reach equity in mathematics and science educational outcomes for all students.

Keywords: sex differences, mathematics, science, education, meta-analysis

The issue of sex differences in science and mathematics achievement continues to capture the interest of parents, educators, researchers and policy-makers, with implications for the ways in which children are educated and encouraged to pursue their chosen careers (Halpern et al., 2007; Hyde & Lindberg, 2007). While significant inroads have been made in recent decades, women continue to be underrepresented in science, technology, engineering and math (STEM)-related fields (Handelsman et al., 2005; National Science Foundation, 2011), despite the fact that more women than men now attend college than men (Alon & Gelbgiser, 2011). Predicted shortfalls for the USA in the number of science graduates for the USA relative to other developing nations carry serious economic and social consequences (President's Council of Advisors on Science and Technology, 2010), and will require broadening the

pool of new entrants into STEM-fields to include more women in order to meet the growing demand. Though the exact causal mechanisms that contribute to sex differences in entering mathematics and science fields are yet to be fully understood (Ceci & Williams, 2011; Hanson, Schaub, & Baker, 1996), many researchers believe that early sex differences in achievement at school shape attitudes towards STEM-fields and self-efficacy beliefs (Halpern et al., 2007; Newcombe et al., 2009; Wai, Lubinski, & Benbow, 2009; Wang, Eccles, & Kenny, 2013). Furthermore, even if they choose not to pursue a STEM-related profession, students entering college and university are increasingly required to have more advanced technical and quantitative skills. For this reason the emergence of sex differences in educational achievement of students is of interest to educational psychologists.

¹ School of Applied Psychology, Griffith University, Queensland, Australia

² Behavioural Basis of Health Program, Griffith Health Institute, Queensland, Australia

Correspondence should be addressed to:

David Reilly. School of Applied Psychology, Griffith University Email : d.reilly@griffith.edu.au

A key component of any strategy to raise the representation of women in STEM-fields is to address gender gaps in mathematics and science outcomes, but the existence and magnitude of these differences is strongly contested (Gallagher & Kaufman, 2005; Halpern et al., 2007; Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 2006; Spelke, 2005; Wai et al., 2009). Much of the empirical research in this area is somewhat dated (e.g. Hyde, Fennema, & Lamon, 1990). Furthermore, as Hedges and Nowell (1995) point out, with few exceptions most empirical studies in this area are subject to selection and sampling biases. Furthermore, as there are interactions between gender and other sociocultural factors (Becker & Hedges, 1988; Frieze, 2014; Hyde & Mertz, 2009; Nowell & Hedges, 1998; Spelke, 2005) these findings do not necessarily generalize well to the *wider* population. Debate about educational issues such as sex-segregated schooling (Halpern, Eliot et al., 2011), or early intervention programs to boost mathematics and science literacy (Hyde & Lindberg, 2007; Newcombe & Frick, 2010) can only be served by timely and accurate empirical research into the nature of sex differences in science and mathematics achievement (Alberts, 2010; Halpern, Beninger, & Straight, 2011). Additionally, if gender gaps are decreasing in response to cultural and educational changes (Auster & Ohm, 2000; Wood & Eagly, 2012), existing research on sex differences in educational achievement for mathematics and science could quickly become dated and require periodic reassessment (Hyde & Mertz, 2009). We describe the findings of prior research on sex differences in these domains, and then extend these findings by reporting a meta-analysis of sex differences in national science and mathematical achievement from the National Assessment of Educational Progress (NAEP) for the years 1990-2011. First, we review the theoretical frameworks that posit the emergence of sex differences in quantitative reasoning.

The issue of sex differences in science and mathematics achievement continues to capture the interest of parents, educators, researchers and policy-makers, with implications for the ways in which children are educated and encouraged to pursue their chosen careers (Halpern et al., 2007; Hyde & Lindberg, 2007). While significant inroads have been made in recent decades, women continue to be underrepresented in science, technology, engineering and math (STEM)-related fields (Handelsman et al., 2005; National Science Foundation, 2011), despite the fact that more women than men now attend college than men (Alon & Gelbgiser, 2011). Predicted shortfalls for the USA in the number of science graduates for the USA relative to other developing nations carry serious economic and social consequences (President's Council of Advisors on Science and Technology, 2010), and will require broadening the pool of new entrants into STEM-fields to include more women in order to meet the growing demand. Though the exact causal mechanisms that contribute to sex differences in entering mathematics and science fields are yet to be fully understood (Ceci & Williams, 2011; Hanson, Schaub, & Baker, 1996), many researchers believe that early sex differences in achievement at school shape attitudes towards STEM-fields and self-

efficacy beliefs (Halpern et al., 2007; Newcombe et al., 2009; Wai, Lubinski, & Benbow, 2009; Wang, Eccles, & Kenny, 2013). Furthermore, even if they choose not to pursue a STEM-related profession, students entering college and university are increasingly required to have more advanced technical and quantitative skills. For this reason the emergence of sex differences in educational achievement of students is of interest to educational psychologists.

A key component of any strategy to raise the representation of women in STEM-fields is to address gender gaps in mathematics and science outcomes, but the existence and magnitude of these differences is strongly contested (Gallagher & Kaufman, 2005; Halpern et al., 2007; Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 2006; Spelke, 2005; Wai et al., 2009). Much of the empirical research in this area is somewhat dated (e.g. Hyde, Fennema, & Lamon, 1990). Furthermore, as Hedges and Nowell (1995) point out, with few exceptions most empirical studies in this area are subject to selection and sampling biases. Furthermore, as there are interactions between gender and other sociocultural factors (Becker & Hedges, 1988; Frieze, 2014; Hyde & Mertz, 2009; Nowell & Hedges, 1998; Spelke, 2005) these findings do not necessarily generalize well to the *wider* population. Debate about educational issues such as sex-segregated schooling (Halpern, Eliot et al., 2011), or early intervention programs to boost mathematics and science literacy (Hyde & Lindberg, 2007; Newcombe & Frick, 2010) can only be served by timely and accurate empirical research into the nature of sex differences in science and mathematics achievement (Alberts, 2010; Halpern, Beninger, & Straight, 2011). Additionally, if gender gaps are decreasing in response to cultural and educational changes (Auster & Ohm, 2000; Wood & Eagly, 2012), existing research on sex differences in educational achievement for mathematics and science could quickly become dated and require periodic reassessment (Hyde & Mertz, 2009). We describe the findings of prior research on sex differences in these domains, and then extend these findings by reporting a meta-analysis of sex differences in national science and mathematical achievement from the National Assessment of Educational Progress (NAEP) for the years 1990-2011. First, we review the theoretical frameworks that posit the emergence of sex differences in quantitative reasoning.

Theoretical Perspectives on Sex Differences in Quantitative Reasoning

While reviews of intelligence testing studies find no evidence for sex differences in general intelligence (Halpern & Lamay, 2000; Neisser et al., 1996), consistent patterns of sex differences have been observed for more *specific* components of cognitive ability (Halpern, 2011; Kimura, 2000). For example, women show greater proficiency with verbal ability and language tasks while men demonstrate higher performance on tasks that tap visuospatial abilities (Halpern & Lamay, 2000). Sex differences have also been documented in quantitative reasoning (our present focus), which include tasks that assess mathematical and scientific

skills (Halpern et al., 2007; Wai et al., 2009). A number of theoretical perspectives have been proposed by researchers to explain why sex differences in quantitative reasoning might emerge; these include both biological and psychosocial contributions. While a full critique of all of these theoretical perspectives is beyond the scope of this study, the most prominent and well-established perspectives may be categorized as biological, social/environmental, or psychobiosocial theories.

Biological Theories of Sex Differences

Sex hormones have been proposed as an explanation for group differences between males and females (Collins & Kimura, 1997; Kimura, 2000), as sex hormones exert an influence on the organization and development of the human brain before birth (Hines, 2006), as well as playing an activational role at different points in maturation (Hines, 1990). Associations have been found between digit ratio - a marker of prenatal androgen exposure - and some cognitive tasks (Collaer, Reimers, & Manning, 2007), though evidence has been mixed. However most research on biological contributions to sex differences has focuses on differences in sex hormone production, which increases with the onset of puberty. Since this also coincides with a widening of the gender gap in quantitative reasoning during adolescence and early adulthood (Hyde et al., 1990), there is an intuitive appeal to such an explanation. While initial interest by researchers into the contributions of sex hormones such as androgens to sex differences in quantitative reasoning was high (Kimura & Hampson, 1994), research findings have found mixed support with some studies finding no association while other studies observing that endogenous hormone levels explain very little variance in individual performance (Halari et al., 2005; Puts et al., 2010).

Another purported biological contribution to sex differences in quantitative reasoning comes from evolutionary psychology. Darwin (1871) first proposed that sexual selection as a result of evolutionary pressures has led to a differentiation in the roles of men and women, a theme that has been expanded upon by evolutionary psychology to propose an alternate explanation for why sex differences in quantitative reasoning emerge (Archer, 1996; Geary, 1996). In the past, it was adaptive for males to develop and hone spatial skills for navigation and hunting (Buss, 1995), leading to the development of greater visuospatial ability in males. This in turn lays down the foundation for the development of quantitative reasoning through a variety of mechanisms including differing social roles and sex-typing of children's' play activities (Caplan & Caplan, 1994; Geary, 1996, 2010). Furthermore, the traditionally feminine roles of caring for others and sensitivity to emotions may have been adaptive, resulting in a tendency for women to focus on people over things (Su, Rounds, & Armstrong, 2009), which Hyde (2014) argued may decrease motivation to acquire quantitative skills and pursue a STEM-based career. A common theme in such arguments is an interaction between biology and environment, rather than a strictly deterministic role of biology.

Social and Environmental Contributions

Although biological factors may make a modest contribution to sex differences, many theorists argue that psychological and social factors exert a greater influence over the course of a life time. One such theory is Eagly and Wood's social-role theory (Eagly, 1987; Eagly & Wood, 1999), which proposes that any psychological sex differences arise from the distribution of men and women's roles in society. The gendered division of labor between men and women encourages the development of instrumental and achievement-oriented traits in men, and expressive and communal-oriented traits in women. Such a position is also compatible with gender schema theory (Bem, 1981), which proposes that children develop an internal schema about the sex-typing of interests and behavior, and that they are motivated to behave in a manner consistent with their internal sex-role identity (Martin & Ruble, 2004). From an early age children learn to categorize things as inherently masculine or feminine (Kagan, 1964), including school subjects like mathematics and science (Nosek et al., 2009). These form the foundation for sex-typing of interests and activities, which facilitate the development of specific cognitive abilities. Nash (1979) formalized this as a sex-role mediation explanation for cognitive sex differences, theorizing that masculine identification leads to cultivation of spatial, mathematical and scientific skills (Reilly & Neumann, 2013; Signorella & Jamison, 1986).

Another prominent theory was put forward by Caplan and Caplan (1994), who argued that traditionally "masculine" play activities promote the development of spatial ability by encouraging the practice and application of spatial skills (Serbin, Zelkowitz, Doyle, Gold, & Wheaton, 1990). Other theorists argue that gender conformity pressures also play an affective role in developing one's talents. Highly sex-typed individuals are motivated to keep their behavior consistent with internalized sex-role standards and norms, while those low in sex-typing show greater cognitive and behavioral flexibility (Bem, 1975; Martin & Ruble, 2004; Spence, 1984). This has implications for success in academic domains that are traditionally male-dominated, such as science and mathematics (Eccles, 2007). Conversely, as we see changes in the segregation of men and women's roles and increasing gender equality, we might also see a diminishing of sex differences in these areas over time (Hyde, 2014).

Psychobiosocial Theories of Sex Differences

While theorists may be divided over the relative share of nature and nurture in the emergence of sex differences in cognitive abilities, there is a growing consensus that both make a meaningful contribution and neither in isolation cannot explain sex differences (Wood & Eagly, 2013). Indeed, it may be impractical to separate a specific biological and social component and study them in isolation, as their effects are reciprocal in nature (Halpern, 2011). Many theorists have therefore adopted psychobiosocial models for explaining the development of sex differences (Halpern &

Tan, 2001; Hausmann, Schoofs, Rosenthal, & Jordan, 2009); these incorporate elements of biological, psychosocial and sociocultural factors to explain group differences between males and females at the population level.

While these theories offer perspectives on why sex differences in quantitative reasoning may be found, it is also important to consider the many ways in which males and females are alike. Hyde (2005) has proposed the *gender similarities hypothesis*, which argues that men and women are more similar than different. Specifically, it hypothesizes that sex differences in cognition are either small in magnitude or nonexistent. While this hypothesis is not supported for language (Lynn & Mikk, 2009; Stoet & Geary, 2013) and spatial abilities (Voyer, Voyer, & Bryden, 1995) where sex differences are moderately large, the gender similarities hypothesis may be compatible with the existence of sex differences in quantitative reasoning, as these tend to be somewhat smaller in magnitude (Hyde et al., 1990). However the gender similarities hypothesis would be incompatible with sex differences that are moderate or large in magnitude, such as a gender imbalance in the sex ratio of high achieving students in mathematics and science (Benbow, 1988; Hedges & Nowell, 1995). It is also a hypothesis that is can easily be put to the test, by examining the performance of men and women in tests that tap quantitative reasoning skills.

Previous Meta-Analyses of Sex Differences in Mathematics and Science

Meta-analysis of national testing data by Hedges and Nowell (1995) from several decades of assessment (1960s – 1990's) revealed small mean differences favoring males in mathematics and science performance (ranging from $d = +.03$ to $d = .26$ for mathematics, and $d = +.11$ to $d = +.50$ for science). Although mean sex differences might play an important role in the underrepresentation of women in STEM-fields, other researchers have noted that the distribution of performance in a number of cognitive domains is more variable for males than females (Feingold, 1992; Hyde, 2005; Machin & Pekkarinen, 2008). Even if there were no differences in the *average* performance of males and females on a specific ability test, greater variance in the male group would result in an overrepresentation in the extreme tails of the distribution (Feingold, 1992; Halpern et al., 2007; Turkheimer & Halpern, 2009), such as the intellectually gifted from which many STEM-researchers hail (Wai, Cacchio, Putallaz, & Makel, 2010). For example, sex (male:female) ratios of students at the 95th percentile in the above-mentioned datasets ranged from 1.5 to 2.4 in mathematics, and 2.5 to 7.0 in science achievement across samples (Hedges & Nowell, 1995). This can translate to a disparity in educational outcomes, and some researchers argue that sex differences in variability may be more important than the mean differences (Feingold, 1995; Humphreys, 1988; Machin & Pekkarinen, 2008).

The greater male variability hypothesis can be examined through calculation of the variance ratio (VR),

defined as the ratio of male variance to female variance (Feingold, 1992; Hedges & Nowell, 1995; Turkheimer & Halpern, 2009). A variability ratio of 1.00 indicates that males and females are equal in variance. VR values less than 1.00 indicate that females show more variability than males, while VR values greater than 1.00 reflect greater male variability (Priess & Hyde, 2010). Feingold (1994) argues that values between .90 and 1.10 ought to be regarded as negligible (i.e. homogeneity of variance), and this practice is adopted herein.

More recently, Hyde et al. (2008) presented data from a subset of the National Assessment of Educational Progress (NAEP), a nationally representative probability sample drawn from all 50 states of the USA. The advantage of this sampling method is that national NAEP data is a reliable population level-estimate of student performance, reflecting the demographic traits of the general population of students. Although individual state and national performance data was not available at the time, Hyde and colleagues obtained data from a selection of ten states across Grades 2 through 11. Mean sex differences were small (d 's from $-.02$ to $.06$). Hyde (2009) has characterized these differences as 'trivial' in size and others have used this research to argue that sex differences are now no longer found in modern samples (Hyde & Mertz, 2009; Lindberg, Hyde, Petersen, & Linn, 2010).

Although the analysis of Hyde et al. (2008) was conducted using the most recent information available at the time, a key limitation of their methodology is that only a ten-state *subset* of the national dataset was analyzed. Hedges and Nowell (1995) argued there are limitations to the use of samples that show a selection bias, because the conclusions they yield may be erroneous if attempting to generalize to the wider population (Becker & Hedges, 1988; Spelke, 2005; Stumpf, 1995). In particular, this may affect the magnitude of any observed gender gap, as literature suggests an interaction between student and socioeconomic background for many cognitive abilities (Hanscombe et al., 2012; Levine, Vasilyeva, Lourenco, Newcombe, & Huttenlocher, 2005). National assessments of the NAEP are also drawn from both public and private schools, and thus may better reflect the demographic composition of students enrolled in USA educational institutions than analysis of only public school data.

The national test data from the NAEP is now publicly available for researchers, providing a broader sampling of students than was available at the time to Hyde et al. (2008). We present an analysis of national NAEP performance for boys and girls, allowing for an empirical test of claims of sex differences in mathematics for USA students in the present day. Furthermore, because data is now available across several decades, it is possible to examine temporal trends across the year of assessment, as well as developmental trends across grade level of students (Hyde et al., 2008; Lindberg et al., 2010). While the NAEP assesses mathematics more regularly, periodic national testing of science performance makes it possible to assess gender gaps in this

domain as well. Sex differences in science achievement may also play a role in the decision of individuals to pursue a science-related profession.

We focused on four key research questions for the domains of mathematics and science. Firstly, are there sex differences in overall mathematics and science achievement for modern samples of students in the USA, and is the gap diminishing over time? Sociocultural theories of sex differences would predict a decline in the magnitude of sex differences over time, while biological and psychobiosocial theories would be compatible with stability in effect sizes.

Secondly, do males show greater variability in performance than females as predicted by biological theories? Thirdly, if there are sex differences in means and in variance, what is their combined contribution to the proportion of males and females attaining an ‘advanced’ proficiency standard in mathematics and science achievement? Finally, if there are sex differences in science achievement, are they present for all of the three content areas assessed (earth science, physical sciences, life sciences)? These research questions also provide a test of the sex differences and similarities hypothesis, which would predict that effect sizes are small in magnitude.

Method

National Assessment of Educational Progress (NAEP) Datasource

The NAEP is a project of the National Center for Education Statistics (NCES), part of the U.S. Department of Education. NAEP conducts assessments across a range of subjects, including reading, writing, mathematics, history, civics, geography and science. Each subject area is assessed periodically, and the most frequently assessed subjects are reading, mathematics, and science. National and state performance in each assessment is reported publicly in a series of documents titled “The Nation’s Report Card” which provides a review of major trends using language accessible to parents, educators, and policy makers (<http://nationsreportcard.gov/>). These form part of the main NAEP assessment, which uses a modern mathematics and science curriculum with large sample sizes and frequent assessments. A secondary category of assessment is the NAEP long term trends (LTT) assessment of mathematics, which samples students using an earlier curriculum framework from the 1970’s onward. The LTT assesses more basic mathematical content, such as numbers, shapes, measurement and probability, while the main assessment also includes algebra, geometry and problem solving. Additionally, the LTT restricts students to hand calculations, which limits the depth of complexity for assessment items. Although useful information can be obtained from the long-term trend assessments, it fails to adequately assess students’ knowledge of more advanced mathematical content included in the main assessment frameworks and is sampled less frequently than the main assessment. As such, it was deemed

unsuitable for analysis, and only the main assessment data was reported in the main article. However published reports of the LTT long-term assessments show a consistent gender gap in favor of males in mathematics for students at age 13 and 17 that has remained essentially unchanged since assessments began (Rampey, Dion, & Donahue, 2009).

The results of NAEP assessments are made freely available to researchers for secondary analysis via the NAEP Data Explorer (<http://nces.ed.gov/nationsreportcard/naepdata/>). The target population for NAEP national assessments is made up of all students in any educational institution (from both private and public schooling), currently enrolled in the target grade (4, 8, and 12). School and student responses are appropriately weighted to draw an estimate of the target population that reflects student demographics (for example, specific ethnic and socioeconomic groups). This may mean that some students and schools will be over-sampled or under-sampled as appropriate. These weights are applied to draw an estimate of national student performance, reported through the NAEP Data Explorer. Additional information about sampling design is available from the NAEP website (<https://nces.ed.gov/nationsreportcard/mathematics/samplede sign.asp>)

Mathematics Framework. The mathematics assessment framework covers five key content areas, which have remained the same since 1990. These are (a) number properties and operations, (b) measurement, (c) geometry, (d) data analysis, statistics, and probability, and (e) algebra. Students are assessed at a grade-level appropriate standard (for example, at grade 8 the topic of algebra includes linear equations, while at grade 12 this is extended to include quadratic and exponential equations). Assessment items vary in complexity level to accommodate a wide range of ability levels, which is important as some research has noted greater sex differences are present for complex problem solving items (Hyde et al., 1990). Calculators are permitted for approximately one third of the assessment, while the remaining questions must be completed without calculators. The mathematics framework for assessment of Grades 4 and 8 is comparable with earlier assessments, allowing student performance in more recent years to be compared to those from earlier assessments. Although a revised mathematics framework was instituted in 2005 for students in Grade 12, these assessments are comparable to those administered previously as they reflect similar content areas. Further information on the mathematics content areas can be found at the NAEP website, <http://nces.ed.gov/nationsreportcard/mathematics/whatmeasure.aspx>

Science Framework. Topic areas for science assessment are grouped into the following three domains, which form separate subscales as well as contributing to the overall science achievement score:

- **Physical sciences**, including concepts related to properties and changes of matter, forms of energy, energy transfer and conservation, position and motion of objects, and forces affecting motion.

- **Life sciences**, including organization and development of cells and organisms, matter and energy transformations, interdependence, heredity and reproduction, evolution and diversity.

- **Earth and space sciences**, including concepts relating to objects in the universe, the history of the Earth, material properties, tectonics and energy in Earth systems, climate and weather, and biogeochemical cycles.

The science framework used for assessment was revised in 2009, in response to revised national science education standards. While the content areas remained the same (physical, earth and life science), they now include coverage of space science. Students completed a range of multiple-choice and open-ended questions, which also include hands-on practical science tasks and interactive computer-administered tasks from the 2009 assessment onward. For additional information about the science framework and sample questions, see <http://nces.ed.gov/nationsreportcard/science/whatmeasure.aspx>

Reliability of NAEP instrument. Multiple choice items are computer scored, while constructed response are marked by raters. Consistency across markers for the constructed response items was generally high for both mathematics and science (Cohen's Kappa > .80). Item response theory (IRT) is then employed by NCES to measure latent scores, which offers greater control over the measurement characteristics of each question and ensures high reliability. For additional information about reliability of measures, see <http://nces.ed.gov/nationsreportcard/tdw/analysis/>. Furthermore, the NCES conducted a NAEP-TIMSS linking study to compare the assessment frameworks to international standards, finding them comparable.

Schedule of Assessment

Mathematics and science assessments are conducted periodically, adhering to the NAEP schedule. Mathematics is assessed more frequently, roughly every two to three years (1990, 1992, 1996, 2000, 2003, 2005, 2007, 2009, 2011). The schedule of assessments gives greater coverage to students in grades 4 and 8, which are developmentally critical time periods for the acquisition of mathematics and scientific skills (Newcombe & Frick, 2010). Grade 12 assessment was not conducted in the years 2003 and 2011. Science is assessed every four to five years (1996, 2000, 2005, 2009, 2011) and recruited somewhat smaller samples of students than the mathematics assessments. Grades 4, 8 and 12 were all assessed in the science target years, except for 2011.

In addition to an overall test score, students are evaluated against fixed achievement levels in the NAEP, categorizing students at a basic, proficient, and advanced level. Sex differences in the percentage of students attaining these levels are also available, and were obtained from the NAEP Data Explorer. While some researchers have examined sex differences in the extreme upper-tail of mathematics and science distributions (Benbow, 1988; Hedges & Nowell, 1995; Hyde et al., 2008; Nowell &

Hedges, 1998; Wai et al., 2010), Hyde and colleagues (2009) have questioned whether sex differences in *extreme* talent are a necessary requirement for pursuing STEM-related fields. When greater male variability is present, this may present an exaggerated picture of sex differences, particularly if more stringent cutoff points are examined (e.g. 99.9th percentile). Examining sex ratios in attainment of an advanced proficiency in science or mathematics represents a tradeoff between selecting a cutoff point that is germane to the question of underrepresentation of women in STEM-related fields, and seeking to avoid selecting an ability level that serves to exaggerate sex differences.

Participants

National performance data in NAEP mathematics was examined for the period 1990 – 2011, with a combined total sample size of almost 2 million students (see Table 1). Performance data in science was examined for the period 1996 – 2011. Science was assessed less frequently, and with fewer students, with a combined total sample size of over 800,000. Information on sample sizes was obtained from annual reports of the NAEP, which in recent years followed the convention of rounding to the nearest hundred. When individual numbers of males and females were not reported, the assumption of equal sample sizes was made. Additional information on the schedule of assessments and sample size of individual assessment years can be found in the Appendix.

Meta-Analytic Procedure

Mean math and science scores and standard deviation for males and females were obtained from the Data Explorer website. The NAEP Data Explorer provides summary statistics (i.e. mean, standard deviation) rounded off to whole numbers which introduces measurement imprecision, but can also export more precise values in Excel format which was the option used in this meta-analysis. The unit of analysis was group differences in performance of males and females at the national level, rather than for individual states. Effect sizes are reported as the mean difference between males and females in standardized units (Cohen, 1988; Hedges, 2008), commonly referred to as Cohen's *d*. By convention, a positive value for *d* indicates higher male performance while a negative value indicates higher female performance (Hyde, 2005).

Comprehensive Meta Analysis (CMA) V2 and Microsoft Excel software were used to calculate the statistics. Meta-analysis typically employs either a fixed-effects or random-effects model for combining study samples. As NAEP assessments span a number of decades recruiting from independent samples, and it was hypothesized that student characteristics may have changed across years of sampling, a random-effects model was chosen (Borenstein, Hedges, Higgins, & Rothstein, 2009). The random effects model gives slightly wider confidence intervals than a fixed-effects model, but gives a more appropriate estimate of how much variability is present across samples (Hunter & Schmidt, 2000; Kelley & Kelley, 2012). The benefit of such an approach is that we can have greater confidence in the

population estimate of sex differences produced, and that it is not the result of inflated Type I error. Using a random effects statistical model also caters for variation in test content and student characteristics over time.

In addition to the calculation of effect size data for each grade level, we investigated whether the year of assessment was a potential moderator using the technique of meta-regression (Kelley & Kelley, 2012). Meta-regression extends a conventional meta-analysis by determining whether a moderating variable accounts for variation in the magnitude of an observed effect (i.e. explains sources of heterogeneity). Based on claims of diminishing gender gaps (e.g. Hyde & Linn, 2006), a negative association with year of assessment was predicted. While it is clear that sex differences in mathematics are smaller than systematic reviews had found in data from the 1960's – 1980's (Hedges & Nowell, 1995), it is not apparent whether such a trend would continue to the point at which males and females would perform equivalently (Caplan & Caplan, 1994), or whether it would plateau. We employed a random effects model (method of moments) for the meta-regression model to test if the year of assessment acted as a moderator (Borenstein et al., 2009; Thompson & Higgins, 2002). Additionally, subgroup analysis for individual grades using a random effects model was performed to examine whether sex differences change as students progress through their schooling, as indicated by previous research (Hyde et al., 1990).

Variance ratios (VR) for individual samples were calculated following the method of Feingold (1992). Estimates of overall male and female variance ratios were combined across years of sampling for each grade level. Some researchers have questioned whether, in combining variance ratios across samples, mean variance ratios may be the most appropriate measure (Katzman & Alliger, 1992), and have advocated the use of medians or log transformed means. These metrics are most appropriate if the direction of variance ratios change across samples (i.e. greater male variability is found in some samples, while greater female variability is found in others). While this was not the case (see Appendix), by convention and for comparability with other studies the log transformed variance ratios were averaged across sample years and then transformed back into the Fisher's variance ratio statistic. This statistic addresses whether males and females differ at the extreme tails of an ability distribution (for example, the top 1% of gifted students) rather than focusing on the performance of the 'average' students in the middle of the distribution (Priess & Hyde, 2010).

Additionally, the percentages of students for each gender who achieved an 'Advanced' proficiency standard were obtained to investigate the combined effect of sex differences in central tendency and variability. Sex ratios, defined as the relative risk ratio (RR) of male to female students, were calculated for mathematics and science performance at the 'Advanced' level of proficiency. This is a somewhat different methodology than has been followed in previous studies, and represents a tradeoff between selecting

a cutoff-point that fairly evaluates high achieving students in their ability to solve STEM problems, and selecting an arbitrarily high cutoff (e.g. 99th percentile) that would serve to exaggerate sex differences.

Results

Two separate meta-analyses were conducted on the NAEP sample for mathematics and science, with population-level estimates of sex differences partitioned by grade level (4, 8, 12). Although statistically significant sex differences favoring males were found in each grade ($p < .001$), emphasis is placed on *effect size* as this gives an indication of the magnitude and practical impact of the observed differences (Hedges, 2008; Hyde, 2005). In a review of meta-analytic theory and practice, Hyde and Grabe (2008, p. 170) recommend a threshold for considering effect sizes in sex differences research *a priori*, and argued that effect sizes smaller than $d = .10$ be considered "trivial" per Hyde's (2005) gender similarities hypothesis. Accordingly we use this threshold herein for considering whether the observed sex differences are practically meaningful. Variance ratios, and the sex ratio of students attaining the advanced level of proficiency are also reported for maths and science. The original data used in this analysis is presented in the Appendix.

NAEP assessment of mathematics

National performance data in mathematics was examined for the period 1990 – 2011 (see Appendix for schedule of assessment years). National sex differences are somewhat larger than those reported by Hyde and colleagues (2008) in their 10 state sample, with a weighted mean effect size of $d = .07$, $Z = 12.07$, $p < .001$. However there was considerable heterogeneity present in the distribution of effect sizes, $Q(23) = 251.57$, $p < .001$, $I^2 = 90.86$ (see Figure 1). In order to better explain variability across assessments, we tested whether grade level and year of assessment were potential moderators.

Grade level as a moderator. Table 2 presents comparisons between males and females in maths across the three grade levels. When effect sizes were partitioned across the three measured age groups using subgroup analysis, there was a statistically significant difference between grade levels, $Q(2) = 23.15$, $p < .001$. While sex differences were extremely small in elementary and early high school, they grew larger in the final year of high school, $d = .10$. The grade 12 effect size is at the threshold of Hyde's (2005) criterion for non-trivial sex differences.

Year of assessment as a moderator. Next we performed a meta-regression analysis to test for a declining gender gap in mathematics over time. Contrary to our hypothesis, there was no significant effect of assessment year, $Z = -.10$, $b = -.0001$, $CI_{95\%} = -.0016$ to $+.0015$, $p = .923$, nor was the interaction between year and grade significant. This is consistent with other studies that reported stability for mean sex differences in mathematics in recent decades rather

than a declining trend (McGraw, Lubinski, & Strutchens, 2006; Rampey et al., 2009).

Variance Ratios. In line with previous research, the variability of boys' performance in mathematics was wider than that of females across each age group (see Table 3), and exceeded Feingold's (1994) threshold for non-trivial variance ratios. These variance ratios were also stable across the time period examined, with no association with year of assessment or grade, $p > .05$.

Gender Gaps in High Achievers for Mathematics. In order to evaluate the combined effect of mean differences and greater male variability, we calculated the ratio of males:females attaining the advanced proficiency standard for mathematics, $RR = 1.51$, $Z = 15.36$, $p < .001$. As there was significant heterogeneity across assessments, $Q(23) = 300.99$, $p < .001$, $I^2 = 92.35$, we calculated risk ratios separately for each grade level using subgroup analysis (see Table 3). There was a statistically significant difference in sex ratios between grades, $Q(2) = 61.74$, $p < .001$. While there was a moderate overrepresentation of high achieving males in Grades 4 and 8, sex ratios increased considerably by Grade 12 to a ratio of 2.13 males to every female student. While these ratios are still smaller than reported from earlier decades (e.g. Benbow, 1988; Hedges & Nowell, 1995), they remain important targets for educational intervention to encourage and foster high achievement.

Additionally, we tested whether there was a decline in the gender gap for high achievers over time, finding a significant interaction between grade and year of assessment, $p < .05$. To investigate, we performed a meta-regression on year of assessment for each grade level. While there was a tendency towards slightly smaller sex ratios for grade 4 students over time, $Z = -4.45$, $b = -.0247$, $CI_{95\%} = -.0355$ to $-.0138$, $p < .001$, there was no association between year of assessment and high achievers in grades 8 ($Z = -.37$, $p = .711$) and 12 ($Z = -1.15$, $p = .249$) indicating stability across the time period examined.

NAEP assessment of science

National performance data in science was examined for the period 1996–2011 (see Appendix for schedule). Overall, the sex difference between males and females was small and comparable to sex differences in mathematics, $d = .11$, $Z = 9.15$, $p < .001$. However there was considerable heterogeneity across assessments, $Q(11) = 328.22$, $p < .001$, $I^2 = 96.33$ (see Figure 2). In order to better explain variability across assessments, we tested whether grade level and year of assessment were potential moderators.

Grade level as a moderator. Using subgroup analysis we partitioned effect sizes across the three grade levels, reducing heterogeneity somewhat. Table 4 presents sex differences in science achievement across each grade level, and shows significant differences favoring males across all grades. While the observed effect sizes were small in magnitude, values for grade 8 and grade 12 exceed Hyde's (2005) criteria for negligible sex differences (both $d = .12$ and $.13$ respectively).

Year of assessment as a moderator. Next we performed a meta-regression analysis to test the effect of assessment year as a potential moderator. Contrary to our hypothesis of a declining gender gap in science over time, there was no significant effect of the year of assessment on the magnitude of sex differences in science ($b = .00$, $CI_{95\%} = -.0039$ to $+.0057$, $Z = .37$, $p = .711$), nor was there an interaction between year and grade.

Variance Ratios. Consistent with previous research, the variability of boys' performance in science was larger than that of girls' (see Table 5). Variance ratios across all grades exceeded Feingold's (1994) criterion for greater male variability, and were comparable to that found for mathematics. These variance ratios were also stable across the time period examined, with no association with year of assessment or interaction with grade, $p > .05$.

Gender Gaps in High Achievers for Science. The influence of greater male variability is most readily apparent when looking at sex ratios for attainment of an advanced proficiency standard in science. We calculated the risk ratio of males:females attaining the advanced proficiency standard for science, $RR = 1.85$, $Z = 12.81$, $p < .001$. As there was significant heterogeneity across assessments, $Q(12) = 83.32$, $p < .001$, $I^2 = 85.63$, we calculated risk ratios separately for each grade level using subgroup analysis (see Table 5). This reduced heterogeneity somewhat. Sex ratios for students in Grade 4 were modest (1.56), but grew wider for older students in grades 8 (1.88) and grade 12 (2.28). There was also a significant difference in science gender gaps between grades, between groups heterogeneity $Q(2) = 9.05$, $p = .011$.

Additionally, we tested whether there was a decline in the gender gap for high achievers over time, or an interaction between grade and year. While there was no significant association with year of assessment overall, $Z = 0.84$, $p = .401$, the interaction was significant $p < .05$, and we examined effects of year for each level of grade. While there was no significant association with year of assessments for grades 4 ($Z = -.13$, $p = .899$) and 12 ($Z = -.58$, $p = .557$), there was a significant trend towards slightly larger science sex ratios in more recent years for students in grade 8, $Z = 2.98$, $b = -.0260$, $CI_{95\%} = -.0009$ to $+.0431$, $p = .003$.

Science Domains. Overall science achievement only shows part of the picture, however. NAEP assesses science literacy across three subject domains: physical sciences, earth sciences, and life sciences (see Table 6). If group differences were present across all *three* domains then sex differences in *overall* science literacy might be an appropriate target for intervention. However, this was not the case. While small sex differences were found in physical ($d = .13$) and earth sciences ($d = .17$), there were no significant differences for the field of life science. The absence of a statistically significant sex difference in life sciences is consistent with the findings of the National Educational Longitudinal Study (Burkam, Lee, & Smerdon, 1997), and the Trends In Mathematics and Science Study (Neuschmidt, Barth, & Hastedt, 2008) which report finding no sex differences in the

field of life sciences. However we note that greater male variability was present for all content areas and grades.

There was also considerable heterogeneity of effect sizes across assessments, which may be due in part to the reduced coverage of assessments conducted for science, as well as the smaller sample sizes employed (particularly for grade 12). Accordingly, moderator analysis was also performed for each science content domain to determine if grade and year effects were present. There was no effect of year of assessment across all three measures, or interactions between grade and year of assessment. While there were no significant effects of grade level for earth and life sciences, there was a tendency for larger sex differences in physical sciences for older students.

Discussion

The aim of this study was to evaluate the evidence for sex differences in mathematics and science achievement over a broad span of years, and to determine whether these were diminishing over time in response to educational advancements and cultural changes in the roles of men and women (Auster & Ohm, 2000; Wood & Eagly, 2012). The NAEP dataset provided an extremely large nationally representative sample of students collected over a wide timespan, and affords a more accurate and reliable test of sex differences in STEM achievement than can be obtained from a single sample. In doing so it extends coverage of the earlier analysis by Hedges and Nowell (1995) to include the most recently available data (1990-2011).

Sex Differences in Means

In contrast to the analysis by Hyde et al. (2008), which found no difference in a 10-state subset of the national assessment, analysis of the complete NAEP dataset found a small but non-trivial mean difference in mathematics favoring males for students in their final year of year of schooling. Furthermore, we extended the analysis to include national testing of science achievement with similar findings. These findings make the claim that sex differences in quantitative reasoning have been eliminated in modern samples somewhat premature, but neither is there evidence of a wide disparity between the performance of the average male and female student. It is also consistent with US performance in international tests of science and mathematics, which have found only small sex differences (Else-Quest, Hyde, & Linn, 2010; Guiso, Monte, Sapienza, & Zingales, 2008; Reilly, 2012).

It is unclear exactly why the earlier meta-analysis by Hyde et al. (2008) on a small subset of testing data found no difference in NAEP mathematics performance, while sex differences in the national dataset were somewhat larger. It may be due to educational factors (inherent differences from state to state), from the inclusion of private and public institutions in the national dataset, or that when a more representative sample and less selective sample is collected greater sex differences emerge (Hyde et al., 1990). We also note that the magnitude of these mean sex differences in the

NAEP was smaller than similar assessments collected in the decades prior to 1990 for mathematics and science (Hedges & Nowell, 1995), which would be consistent with changes predicted by sociocultural perspectives. However, there was no association between the magnitude of the sex difference observed in each assessment and the assessment year, indicating that there was stability across the period of time investigated (1990-2011). That no further change occurred over this timeframe would be compatible with biological and psychobiological perspectives. Stability is also consistent with the findings of McGraw et al., (2006), who found no change across a shorter timeframe for NAEP mathematics performance. While we found meaningful sex differences, this does not necessarily preclude Hyde's gender similarities hypothesis as it posits that sex differences in cognitive ability are only small in magnitude.

The data also indicated that there was a developmental trend across both types of quantitative reasoning skills, with smaller effect sizes in elementary school and larger effect sizes in older students. Sex differences in mathematics exceed Hyde's criterion in Grade 12, while sex differences in science achievement reach a non-trivial size in Grades 8 and 12. A prior meta-analysis by Hyde et al. (Hyde et al., 1990) also found larger sex differences are observed when complex problem solving tasks are measured, and the mathematics assessment framework increases in complexity during grades 8 and 12. This is also consistent with developmental literature reporting a widening of the gender gap in quantitative reasoning at around puberty and middle-school (Fan, Chen, & Matsumoto, 1997; Hyde et al., 1990; Robinson & Lubinski, 2011), when the saliency of gender roles becomes more prominent as suggested by sociocultural perspectives on gender (Nash, 1979; Ruble, Martin, & Berenbaum, 2006). During adolescence and into early adulthood, gender stereotyping about the sex-typing of activities and interests increases at both the explicit and implicit level (Halpern & Tan, 2001; Nosek et al., 2009; Steffens & Jelenec, 2011), which has implications for sex differences in achievement motivation and self-efficacy for mathematics and science (Priess & Hyde, 2010; Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). However it also coincides with a time of increased hormonal changes as outlined by biological theories (Kimura, 2000), and therefore it is difficult to offer more than speculation as to the origins of sex differences at these developmental periods.

Of particular interest in our analysis is the observation that mean sex differences were present for some, but not all, of the scientific domains assessed by the NAEP. Despite the considerable sample size there was no sex difference found for biology and life sciences, where males and females show equivalent performance (Neuschmidt et al., 2008). Reviews of the literature find that males have greater overall interest in science than females and rate their aptitude more highly (Osborne, Simon, & Collins, 2003; Weinburgh, 1995), but that when inquiries are made regarding interest in specific scientific domains, biology and life sciences show no significant difference between males and females (Miller, Blessing, & Schwartz, 2006). Rather than indicating any

inherent lack of ability, sex differences in certain but not all domains of science may reflect different patterns of interest and motivation towards people-oriented fields (Su et al., 2009), or that other domains are seen as being less relevant to future career paths (Jones, Howe, & Rua, 2000; Miller et al., 2006). Alternately, the mathematical requirements of biology and life sciences may be lower than for the physical sciences or there may be reduced sex-typing stereotypes for this field of study.

High Achievers

While sex difference research often focuses on the performance of the *average* student, considerably less attention is given to sex differences in the prevalence of high achievers and those factors that contribute to their success (Wai, Putallaz, & Makel, 2012). Although only small mean differences in mathematics and science achievement were found, consistent with prior research the performance of males showed consistently greater variability than that of female students (Hedges & Nowell, 1995). Greater male variability in performance is often associated with essentialist biological theories of sex differences (Feingold, 1992), but it is also predicted by differential social and learning experiences afforded to boys and girls argued in sociocultural theories of gender. The combined effect of small mean differences and greater male variability is then reflected in the sex ratios of students attaining the “high” proficiency standard of the NAEP in maths and science. While there are no established guidelines as to how to interpret the magnitude of sex ratios, we would suggest that a sex ratio of over 2:1 (i.e. over twice as many males as females reaching these standards) should be considered meaningful and nontrivial. Finding a large sex difference in high achievers for mathematics and science may not be in keeping with a strict interpretation of Hyde’s (2005) gender similarities hypothesis, but it should be noted that the hypothesis as it was *originally* articulated considered only mean sex differences (Hyde, 2005), and did not speak to gender imbalances in high achievers. Additionally, there was no overall effect of year of assessment on tail ratios, though there was a slight tendency for change in grade 4 mathematics and grade 8 science. It may be the case that changes predicted by sociocultural perspectives operate over a longer timeframe, or that greater male variability remains unchanged as might be predicted by psychobiological theories.

Implications

Although mean sex differences in mathematics and science were only small in magnitude, even small differences in ability level may be consequential if experienced over time (Eagly, Wood, & Diekmann, 2000; Prentice & Miller, 1992; Rosenthal, 1986). In particular, they may serve to undermine self-efficacy and interest in traditionally sex-typed subjects such as mathematics and science (Eccles, 2013; Else-Quest, Mineo, & Higgins, 2013). However this is less concerning than the combined effect of small mean differences and greater male variability, which leads to large gender gaps in high-achievers for mathematics and science.

Further efforts may be warranted to encourage and cultivate girls’ interest and aptitude in these subject areas – particularly with students who have yet to realize their full potential. Many students have a stereotypically masculine image of mathematics and science (Nosek, Banaji, & Greenwald, 2002; Smeding, 2012) and countering deeply ingrained sex-stereotypes is not easily achieved (Shapiro & Williams, 2012). While all students receive instruction in these areas through the school curriculum, parents can facilitate development of mathematics and science interest and aptitude by providing early enrichment activities and science learning experiences equally for daughters and sons (Newcombe & Frick, 2010). Boys report having more extracurricular experiences with toys and games that promote science learning (Jones et al., 2000), and examination of parent-child interactions shows that parents explain scientific concepts to boys more frequently than to girls (Crowley, Callanan, Tenenbaum, & Allen, 2001; Diamond, 1994; Tenenbaum & Leaper, 2003). Parents also estimate the intelligence of sons as being higher than that of daughters, including mathematics intelligence (Furnham, Reeves, & Budhani, 2002), and parental expectations can profoundly impact the self-efficacy of children (Eccles, Jacobs, & Harold, 1990). Encouraging and supporting daughters who show interest or aptitude in science to develop their potential may be critical for addressing gender gaps in high-achievers.

The educational environment in which mathematics and science are taught at school can also have a profound impact on student learning outcomes (Gunderson, Ramirez, Levine, & Beilock, 2012). Teachers have different beliefs about male and female students in mathematics, have more frequent interactions with male students than with female, and higher expectations in this field for boys (Li, 1999). Similar findings have been reported for science education, such as calling more frequently on male students to answer questions or provide a demonstration (Jones & Wheatley, 1990). Differential learning experiences for boys and girls in the classroom are often subtle (Beaman, Wheldall, & Kemp, 2006), but may be contributing to the development of lower self-efficacy and less interest in STEM for girls (for a review see Gunderson et al., 2012). Individual differences in endorsement of sex-stereotypes about STEM can seriously undermine girls’ achievement in these fields later in life (Schmader, Johns, & Barquissau, 2004), so it is important that educators send a positive message about the applicability of mathematics and science skills to *both* genders.

A growing body of research also suggests that visuospatial skills play an important role in the development of quantitative reasoning (Nuttall, Casey, & Pezaris, 2005), and that sex differences in spatial ability may be a mediator (Wai et al., 2009). However even brief educational interventions can show marked improvements in the development of spatial ability in both genders (Uttal et al., 2013), with evidence of transfer to other quantitative tasks. Many researchers have advocated for the inclusion of spatial learning within the school curriculum (Newcombe & Frick, 2010; Priess & Hyde, 2010), as this would provide benefits to *all* students and lay down a solid foundation for the later

development of quantitative reasoning. Contrary to our hypothesis, mean sex differences and sex ratios of high achievers did not show a decline over the time period analyzed. Despite societal changes in the roles of men and women (Auster & Ohm, 2000) this has not translated into diminishing sex differences over time as predicted by social and psychobiosocial perspectives. The present findings of stable sex differences give further weight to arguments that educational interventions are still required in the interest of gender equity.

Strengths and Limitations

The issue of sex differences in quantitative reasoning has been contentious in recent decades, with some researchers arguing that there are considerable differences and others that there are none. By employing a large nationally representative sample such as the NAEP, we can be more confident that the observed sex differences reflects the diversity of socioeconomic status and ethnicity found in the United States, as well as the different educational environments of each state. The statistical technique of meta-analysis makes it possible to aggregate findings from multiple waves of assessment, ensuring that the conclusion reached is not idiosyncratic to a particular assessment year and student cohort. As such it gives greater confidence in estimating the magnitude of sex differences in mathematics and science in USA students under the NAEP.

It has also offered the opportunity to test whether the magnitude of said differences is declining, and to establish that – at least for the time period analyzed – these are stable across time. It also draws attention to the role that greater male variability can play, and the critical importance of examining tail ratios of high achieving students for a complete test of the gender similarities hypothesis.

While adding to the existing literature on sex differences, this study is not without limitations. Firstly, it does not provide any information on the causal factors that explain *why* sex differences emerge. Although researchers have identified a number of biological, psychological and social factors that contribute to sex differences in quantitative reasoning (Halpern et al., 2007), many researchers agree that a variety of factors are ultimately responsible and advocate a biopsychosocial model of sex differences (Halpern, 2004; Halpern & Tan, 2001). Thus the findings of a meta-analysis can shed no light on *why* sex differences emerge, and can only document their existence.

Secondly our study does not consider other factors, such as socioeconomic background and ethnicity. There is some evidence to show interactions between sex differences and ethnic backgrounds. For example, while sex differences are consistently found for Caucasian and Hispanic students, some studies have failed to find differences for African American samples (Fan, Chen, & Matsumoto, 1997; McGraw et al., 2006). Likewise some studies have found interactions between socioeconomic status and sex differences in early spatial development (Levine et al., 2005), which provides a foundation for quantitative reasoning. Teasing apart such

theoretical contributions would be a useful addition to the literature. Finally, our analysis is limited by the test content being assessed by the NAEP. Previous studies (e.g. Hyde et al., 1990) have noted larger sex differences are found in complex problem solving, but the NAEP includes test items across a range of difficulty levels. International assessments of student ability such as the Programme for International Student Assessment (PISA) include more challenging test content, and find somewhat larger sex differences in mathematics and science for USA students than under the NAEP (Guiso et al., 2008; Reilly, 2012). While these parallel lines of evidence provide a replication of sex differences, they do suggest that the NAEP may underestimate the true effect size of such differences somewhat.

Summary

In the present study, we report a meta-analysis of sex differences in mathematics and science achievement in the NAEP, a nationally-representative sample of students drawn from public and private institutions from across all states in the USA. Small mean sex differences favoring males were observed in science and mathematics performance making claims of their absence premature. Further examination of male and female performance across the three domains of science found that males and females were equivalent in performance for life sciences, but not for earth and physical sciences. Contrary to our hypothesis, sex differences were not moderated by the year in which students were tested, indicating stability across time. Additionally we found that the performance of males was more variable than that of females, which has implications for the proportion of males to females in the upper-right tail of the ability distribution. Greater male variability may contribute to the disparity in educational outcomes in STEM-related fields with males being over-represented in attainment of an advanced proficiency in mathematics and science by a ratio of over 2:1. Further research into the psychological and social factors underpinning these gender gaps is required, as well as educational interventions and support services to help girls realize their full potential in mathematics and science achievement. Counteracting the tendency for initially small sex differences in achievement to be translated into larger sex differences in career choices is likely to require concerted and sustained efforts at many levels.

References

- Alberts, B. (2010). Policy-making needs science. *Science*, 330(6009), 1287. doi: 10.1126/science.1200613
- Alon, S., & Gelbgiser, D. (2011). The female advantage in college academic achievements and horizontal sex segregation. *Social Science Research*, 40(1), 107-119. doi: 10.1016/j.ssresearch.2010.06.007
- Archer, J. (1996). Sex differences in social behavior: Are the social role and evolutionary explanations compatible? *American Psychologist*, 51(9), 909-917.
- Auster, C. J., & Ohm, S. C. (2000). Masculinity and femininity in contemporary American society: A reevaluation using the Bem Sex-Role Inventory. *Sex Roles*, 43(7/8), 499-528. doi: 10.1023/A:1007119516728

- Beaman, R., Wheldall, K., & Kemp, C. (2006). Differential teacher attention to boys and girls in the classroom. *Educational Review*, 58(3), 339-366. doi: 10.1080/00131910600748406
- Becker, B. J., & Hedges, L. V. (1988). The effects of selection and variability in studies of gender differences. *Behavioral and Brain Sciences*, 11(2), 183-184. doi: 10.1017/S0140525X00049256
- Bem, S. L. (1981). Gender Schema Theory: A cognitive account of sex typing. *Psychological Review*, 88(4), 354-364. doi: 10.1037/0033-295X.88.4.354
- Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences*, 11(2), 169-232. doi: 10.1017/S0140525X00049670
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, UK: Wiley.
- Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal*, 34(2), 297-331. doi: 10.3102/00028312034002297
- Buss, D. M. (1995). Psychological sex differences: Origins through sexual selection. *American Psychologist*, 50(3), 164-168. doi: 10.1037//0003-066X.50.3.164
- Caplan, P. J., & Caplan, J. B. (1994). *Thinking critically about research on sex and gender*. New York: Harper Collins.
- Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 108(8), 3157-3162. doi: 10.1073/pnas.1014871108
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Collins, D. W., & Kimura, D. (1997). A large sex difference on a two-dimensional mental rotation task. *Behavioral Neuroscience*, 111(4), 845-849. doi: 10.1037/0735-7044.111.4.845
- Crowley, K., Callanan, M. A., Tenenbaum, H. R., & Allen, E. (2001). Parents explain more often to boys than to girls during shared scientific thinking. *Psychological Science*, 12(3), 258-261. doi: 10.1111/1467-9280.00347
- Darwin, C. (1871). *The decent of man, and selection in relation to sex*. London: John Murray.
- Diamond, J. (1994). Sex differences in science museums: A review. *Curator: The Museum Journal*, 37(1), 17-24. doi: 10.1111/j.2151-6952.1994.tb01003.x
- Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Hillsdale: Erlbaum.
- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, 54(6), 408-423. doi: 10.1037/0003-066X.54.6.408
- Eccles, J. S. (2007). Where are all the women? Gender differences in participation in physical science and engineering. In S. J. Ceci (Ed.), *Why aren't more women in science? Top researchers debate the evidence* (pp. 199-210). Washington, D.C.: American Psychological Association.
- Eccles, J. S. (2013). Gender and achievement choices. In E. T. Gershoff, R. S. Mistry & D. Crosby (Eds.), *Societal Contexts of Child Development: Pathways of Influence and Implications for Practice and Policy* (pp. 19-34). New York: Oxford University Press.
- Eccles, J. S., Jacobs, J. E., & Harold, R. D. (1990). Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues*, 46(2), 183-201. doi: 10.1111/j.1540-4560.1990.tb01929.x
- 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), 103-127. doi: 10.1037/a0018053
- Else-Quest, N. M., Mineo, C. C., & Higgins, A. (2013). Math and science attitudes and achievement at the intersection of gender and ethnicity. *Psychology of Women Quarterly*, 37(3), 293-309. doi: 10.1177/0361684313480694
- Fan, X., Chen, M., & Matsumoto, A. R. (1997). Gender differences in mathematics achievement: Findings from the National Education Longitudinal Study of 1988. *The Journal of Experimental Education*, 65(3), 229-242. doi: 10.1080/00220973.1997.9943456
- 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. doi: 10.3102/00346543062001061
- Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles*, 30(1), 81-92. doi: 10.1007/BF01420741
- Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist*, 50(1), 5-13. doi: 10.1037//0003-066X.50.1.5
- Furnham, A., Reeves, E., & Budhani, S. (2002). Parents think their sons are brighter than their daughters: Sex differences in parental self-estimations and estimations of their children's multiple intelligences. *The Journal of Genetic Psychology*, 163(1), 24-39. doi: 10.1080/00221320209597966
- Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, 19(2), 229-247. doi: 10.1017/S0140525X00042400
- Geary, D. C. (2010). *Male, female : the evolution of human sex differences* (2nd ed.). Washington, DC: American Psychological Association.
- Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, 320, 1164-1165. doi: 10.1126/science.1154094

- Gunderson, E., Ramirez, G., Levine, S. C., & Beilock, S. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles*, 66(3), 153-166. doi: 10.1007/s11199-011-9996-2
- Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science*, 13(4), 135-139. doi: 10.1111/j.0963-7214.2004.00292.x
- Halpern, D. F. (2011). *Sex differences in cognitive abilities* (4th ed.). Mahwah, NJ: Erlbaum.
- 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. doi: 10.1111/j.1529-1006.2007.00032.x
- Halpern, D. F., Beninger, A. S., & Straight, C. A. (2011). Sex differences in intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge Handbook of Intelligence* (pp. 253-272). New York: Cambridge University Press.
- Halpern, D. F., Eliot, L., Bigler, R. S., Fabes, R. A., Hanish, L. D., Hyde, J. S., et al. (2011). The pseudoscience of single-sex schooling. *Science*, 333(6050), 1706-1707. doi: 10.1126/science.1205031
- Halpern, D. F., & Lamay, M. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review*, 12(2), 229-246. doi: 10.1023/A:1009027516424
- Halpern, D. F., & Tan, U. (2001). Stereotypes and steroids: Using a psychobiosocial model to understand cognitive sex differences. *Brain and Cognition*, 45(3), 392-414. doi: 10.1006/brcg.2001.1287
- Handelsman, J., Cantor, N., Carnes, M., Denton, D., Fine, E., Grosz, B., et al. (2005). More women in science. *Science*, 309(5738), 1190-1191. doi: 10.1126/science.1113252
- Hanscombe, K. B., Trzaskowski, M., Haworth, C. M. A., Davis, O. S. P., Dale, P. S., & Plomin, R. (2012). Socioeconomic status (SES) and children's intelligence (IQ): In a UK-representative sample SES moderates the environmental, not genetic, effect on IQ. *PLoS ONE*, 7(2), e30320. doi: 10.1371/journal.pone.0030320
- Hanson, S. L., Schaub, M., & Baker, D. P. (1996). Gender stratification in the science pipeline: A comparative analysis of seven countries. *Gender and Society*, 10(3), 271-290.
- Hausmann, M., Schoofs, D., Rosenthal, H. E. S., & Jordan, K. (2009). Interactive effects of sex hormones and gender stereotypes on cognitive sex differences--A psychobiosocial approach. *Psychoneuroendocrinology*, 34(3), 389-401.
- Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, 2(3), 167-171. doi: 10.1111/j.1750-8606.2008.00060.x
- Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269(5220), 41-45. doi: 10.1126/science.7604277
- Hines, M. (1990). Gonadal hormones and human cognitive development. In J. Balthazart (Ed.), *Hormones, brain and behaviour in vertebrates* (pp. 51-63). New York: Karger.
- Hines, M. (2006). Prenatal testosterone and gender-related behaviour. *European Journal of Endocrinology*, 155, S115-S121. doi: 10.1530/eje.1.02236
- Humphreys, L. G. (1988). Sex differences in variability may be more important than sex differences in means. *Behavioral and Brain Sciences*, 11(02), 195-196.
- Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, 8(4), 275-292. doi: 10.1111/1468-2389.00156
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60(6), 581-592. doi: 10.1037/0003-066X.60.6.581
- Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, 65(1), 373-398. doi: 10.1146/annurev-psych-010213-115057
- Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, 107(2), 139-155. doi: 10.1037/0033-2909.107.2.139
- Hyde, J. S., & Grabe, S. (2008). Meta-analysis in the psychology of women. In F. Denmark & M. A. Paludi (Eds.), *Psychology of women: A handbook of issues and theories* (pp. 142-173). Westport, CT: Praeger Publishers.
- Hyde, J. S., & Lindberg, S. M. (2007). Facts and assumptions about the nature of gender differences and the implications for gender equity. In S. S. Klein (Ed.), *Handbook for achieving gender equity through education* (2nd ed., pp. 19-32). Mahwah, NJ: Lawrence Erlbaum Associates.
- 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. doi: 10.1126/science.1160364
- Hyde, J. S., & Linn, M. C. (2006). Gender similarities in mathematics and science. *Science*, 314(5799), 599-600. doi: 10.1126/science.1132154
- Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, 106(22), 8801-8807. doi: 10.1073/pnas.0901265106
- Jones, M. G., Howe, A., & Rua, M. J. (2000). Gender differences in students' experiences, interests, and attitudes toward science and scientists. *Science Education*, 84(2), 180-192. doi: 10.1002/(SICI)1098-237X(200003)84:2<180::AID-SCE3>3.0.CO;2-X
- Jones, M. G., & Wheatley, J. (1990). Gender differences in teacher-student interactions in science classrooms. *Journal of Research in Science Teaching*, 27(9), 861-874. doi: 10.1002/tea.3660270906

- Kagan, J. (1964). The child's sex role classification of school objects. *Child Development*, 35(4), 1051-1056. doi: 10.2307/1126852
- Katzman, S., & Alliger, G. M. (1992). Averaging untransformed variance ratios can be misleading: A comment on Feingold. *Review of Educational Research*, 62(4), 427-428. doi: 10.3102/00346543062004427
- Kelley, G., & Kelley, K. (2012). Statistical models for meta-analysis: A brief tutorial. *World Journal of Methodology*, 2(4), 27-32. doi: 10.5662/wjm.v2.i4.27
- Kimura, D. (2000). *Sex and cognition*. Cambridge, MA: MIT Press.
- Kimura, D., & Hampson, E. (1994). Cognitive pattern in men and women is influenced by fluctuations in sex hormones. *Current Directions in Psychological Science*, 3(2), 57-61. doi: 10.1111/1467-8721.ep10769964
- Levine, S. C., Vasilyeva, M., Lourenco, S. F., Newcombe, N. S., & Huttenlocher, J. (2005). Socioeconomic status modifies the sex difference in spatial skill. *Psychological Science*, 16(11), 841-845. doi: 10.1111/j.1467-9280.2005.01623.x
- Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: a review. *Educational Research*, 41(1), 63-76. doi: 10.1080/0013188990410106
- 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), 1123-1135. doi: 10.1037/a0021276
- Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *Trames*, 13(1), 3-13. doi: 10.3176/tr.2009.1.01
- Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science*, 322(5906), 1331-1332. doi: 10.1126/science.1162573
- Martin, C. L., & Ruble, D. N. (2004). Children's search for gender cues : Cognitive perspectives on gender development. *Current Directions in Psychological Science*, 13(2), 67-70. doi: 10.1111/j.0963-7214.2004.00276.x
- McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education*, 37(2), 129-150. doi: 10.2307/30034845
- Miller, P. H., Blessing, J. S., & Schwartz, S. (2006). Gender differences in high school students' views about science. *International Journal of Science Education*, 28(4). doi: 10.1080/09500690500277664
- Nash, S. C. (1979). Sex role as a mediator of intellectual functioning. In M. A. Wittig & A. C. Petersen (Eds.), *Sex-related differences in cognitive functioning: Developmental issues* (pp. 263-302). New York: Academic Press.
- National Science Foundation. (2011). *Women, minorities, and persons with disabilities in science and engineering: 2011*. Arlington, VA: National Science Foundation, Retrieved from www.nsf.gov/statistics/wmpd/pdf/wmpd2011.pdf.
- Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995-2003). *Studies in Educational Evaluation*, 34(2), 56-72. doi: 10.1016/j.stueduc.2008.04.002
- Newcombe, N. S., Ambady, N., Eccles, J. S., Gomez, L., Klahr, D., Linn, M., et al. (2009). Psychology's role in mathematics and science education. *American Psychologist*, 64(6), 538-550. doi: 10.1037/a0014813
- Newcombe, N. S., & Frick, A. (2010). Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education*, 4(3), 102-111. doi: 10.1111/j.1751-228X.2010.01089.x
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math= male, me= female, therefore math≠ me. *Journal of Personality and Social Psychology*, 83(1), 44-59. doi: 10.1037//0022-3514.83.1.44
- Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., et al. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26), 10593-10597. doi: 10.1073/pnas.0809921106
- Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles*, 39(1), 21-43. doi: 10.1023/A:1018873615316
- Nuttall, R. L., Casey, M. B., & Pezaris, E. (2005). Spatial ability as a mediator of gender differences on mathematics tests: A biological-environmental framework. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 121-142). Cambridge, UK: Cambridge University Press.
- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: a review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049-1079. doi: 10.1080/0950069032000032199
- President's Council of Advisors on Science and Technology. (2010). *Prepare and Inspire: K-12 Science, Technology, Engineering, and Math (STEM) Education for America's Future*. Washington, D.C.
- Priess, H. A., & Hyde, J. S. (2010). Gender and academic abilities and preferences. In J. C. Chrisler & D. R. McCreary (Eds.), *Handbook of Gender Research in Psychology* (pp. 297-316). New York: Springer.
- Ramsey, B. D., Dion, G. S., & Donahue, P. L. (2009). *NAEP 2008 Trends in Academic Progress (NCES 2009-479)*. Washington, DC: National Center for Education Statistics.
- Reilly, D. (2012). Gender, culture and sex-typed cognitive abilities. *PLoS ONE*, 7(7), e39904. doi: 10.1371/journal.pone.0039904
- Reilly, D., & Neumann, D. L. (2013). Gender-role differences in spatial ability: A meta-analytic review. *Sex Roles*, 68(9), 521-535. doi: 10.1007/s11199-013-0269-0

- Ruble, D. N., Martin, C. L., & Berenbaum, S. A. (2006). Gender development. In N. Eisenberg, W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (6th ed., Vol. 3. Social, emotional, and personality development, pp. 858-932). Hoboken, NJ: Wiley.
- Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles*, 50(11), 835-850. doi: 10.1023/B:SERS.0000029101.74557.a0
- Serbin, L. A., Zerkowicz, P., Doyle, A. B., Gold, D., & Wheaton, B. (1990). The socialization of sex-differentiated skills and academic performance: A mediational model. *Sex Roles*, 23(11), 613-628. doi: 10.1007/BF00289251
- Shapiro, J., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, 66(3-4), 175-183. doi: 10.1007/s11199-011-0051-0
- Signorella, M. L., & Jamison, W. (1986). Masculinity, femininity, androgyny, and cognitive performance: A meta-analysis. *Psychological Bulletin*, 100(2), 207-228. doi: 10.1037/0033-2909.100.2.207
- Smeding, A. (2012). Women in Science, Technology, Engineering, and Mathematics (STEM): An investigation of their implicit gender stereotypes and stereotypes' connectedness to math performance. *Sex Roles*, 67(11-12), 617-629. doi: 10.1007/s11199-012-0209-4
- Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. *American Psychologist*, 60(9), 950-958. doi: 10.1037/0003-066X.60.9.950
- Steffens, M., & Jelenec, P. (2011). Separating implicit gender stereotypes regarding math and language: Implicit ability stereotypes are self-serving for boys and men, but not for girls and women. *Sex Roles*, 64(5), 324-335. doi: 10.1007/s11199-010-9924-x
- Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PLoS ONE*, 8(3), e57988. doi: 10.1371/journal.pone.0057988
- Stumpf, H. (1995). Gender differences in performance on tests of cognitive abilities: Experimental design issues and empirical results. *Learning and Individual Differences*, 7(4), 275-287. doi: 10.1016/1041-6080(95)90002-0
- Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: a meta-analysis of sex differences in interests. *Psychological Bulletin*, 135(6), 859-884. doi: 10.1037/a0017364
- Tenenbaum, H. R., & Leaper, C. (2003). Parent-child conversations about science: The socialization of gender inequities? *Developmental Psychology*, 39(1), 34-47. doi: 10.1037/0012-1649.39.1.34
- Thompson, S. G., & Higgins, J. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21(11), 1559-1573. doi: 10.1002/sim.1187
- Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., et al. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, 139(2), 352-402. doi: 10.1037/a0028446
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, 117(2), 250-270. doi: 10.1037//0033-2909.117.2.250
- Wai, J., Cacchio, M., Putallaz, M., & Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30 year examination. *Intelligence*, 38(4), 412-423. doi: 10.1016/j.intell.2010.04.006
- 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), 817-835. doi: 10.1037/a0016127
- Wai, J., Putallaz, M., & Makel, M. C. (2012). Studying intellectual outliers: Are there sex differences, and are the smart getting smarter? *Current Directions in Psychological Science*, 21(6), 382-390. doi: 10.1177/0963721412455052
- 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. doi: 10.1177/0956797612458937
- Weinburgh, M. (1995). Gender differences in student attitudes toward science: A meta-analysis of the literature from 1970 to 1991. *Journal of Research in Science Teaching*, 32(4), 387-398. doi: 10.1002/tea.3660320407
- Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In J. M. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. 46, pp. 55-124). New York: Academic Press.
- Wood, W., & Eagly, A. H. (2013). Biology or culture alone cannot account for human sex differences and similarities. *Psychological Inquiry*, 24(3), 241-247. doi: 10.1080/1047840x.2013.815034

Table 1
Sample Size Information for Mathematics and Science Assessments.

| Content Domain | Grade | <i>N</i> of Students Assessed |
|----------------|-------|-------------------------------|
| Mathematics | 4 | 974,700 |
| | 8 | 845,400 |
| | 12 | 104,900 |
| | Total | 1,925,100 |
| Science | 4 | 352,105 |
| | 8 | 470,374 |
| | 12 | 56,437 |
| | Total | 878,916 |

Table 2
Sex differences in NAEP mathematics achievement for Grades 4, 8, 12

| Grade | <i>k</i> | Cohen's <i>d</i> | 95% Confidence Interval | | Test of null (2-tail) | | Heterogeneity |
|-------|----------|------------------|-------------------------|-------------|-----------------------|---------|---|
| | | | Lower limit | Upper limit | Z-value | P-value | |
| 4 | 9 | .07 | .06 | .09 | 10.67 | < .001 | Q(8) = 90.37, $p < .001$, $I^2 = 91.15$ |
| 8 | 9 | .04 | .03 | .06 | 6.54 | < .001 | Q(8) = 28.52, $p < .001$, $I^2 = 71.95$ |
| 12 | 6 | .10 | .08 | .12 | 10.08 | < .001 | Q(5) = 10.71, $p = n.s.$ |

Note: *k* denotes the number of assessments conducted for each grade. Effect sizes that exceed Hyde's (2005) criterion for non-trivial differences ($d \geq .10$) are highlighted in bold.

Table 3

Sex Differences in Variability, and Sex Ratios Attaining Advanced Proficiency in Mathematics

| Grade | Variance Ratio (VR) | Risk Ratio | 95% Confidence Interval | | Test of null (2-tail) | | Heterogeneity |
|-------|---------------------|------------|-------------------------|-------------|-----------------------|---------|--|
| | | | Lower Limit | Upper Limit | Z-value | P-value | |
| 4 | 1.12 | 1.51 | 1.42 | 1.60 | 13.71 | < .001 | Q(8) = 94.30, $p < .001$, $I^2 = 91.51$ |
| 8 | 1.12 | 1.30 | 1.23 | 1.37 | 9.27 | < .001 | Q(8) = 24.71, $p = .002$, $I^2 = 67.63$ |
| 12 | 1.15 | 2.13 | 1.90 | 2.38 | 13.28 | < .001 | Q(5) = 8.77, $p = n.s.$ |

Table 4

Sex differences in NAEP science achievement for Grades 4, 8, 12

| Grade | k | Cohen's d | 95% Confidence Interval | | Test of null (2-tail) | | Heterogeneity |
|-------|-----|-------------|-------------------------|-------------|-----------------------|---------|---|
| | | | Lower Limit | Upper limit | Z-value | P-value | |
| 4 | 4 | .08 | .04 | .12 | 3.64 | .001 | Q(3) = 174.57, $p < .001$, $I^2 = 98.28$ |
| 8 | 4 | .12 | .08 | .16 | 6.39 | < .001 | Q(3) = 41.93, $p < .001$, $I^2 = 90.46$ |
| 12 | 4 | .13 | .09 | .18 | 6.05 | < .001 | Q(3) = 24.68, $p < .001$, $I^2 = 87.84$ |

Note: Effect sizes that exceed Hyde's criterion for non-trivial differences ($d \geq .10$) are highlighted in bold.

Table 5

Sex Differences in Variability, and Sex Ratios Attaining Advanced Proficiency in Science

| Grade | Variance Ratio (VR) | Risk Ratio | 95% Confidence Interval | | Test of null (2-tail) | | Heterogeneity |
|-------|------------------------|---------------|----------------------------|----------------|-----------------------|---------|--|
| | | | Lower Limit | Upper Limit | Z-value | P-value | |
| 4 | 1.09 | 1.56 | 1.33 | 1.83 | 5.45 | < .001 | $Q(3) = 19.20, p < .001,$ $I^2 = 84.37$ |
| 8 | 1.12 | 1.88 | 1.64 | 2.16 | 8.95 | < .001 | $Q(4) = 34.32, p < .001,$ $I^2 = 88.34$ |
| 12 | 1.14 | 2.28 | 1.88 | 2.76 | 8.41 | < .001 | $Q(3) = 4.77, p = n.s.$ |

Table 6
Sex Differences Across NAEP Science Domains for Grades 4, 8 and 12

| Science domain | Grade | Variance Ratio (V/R) | Cohen's <i>d</i> | 95% Confidence Interval | | Test of null (2-tail) | | Heterogeneity |
|------------------|---------|----------------------|------------------|-------------------------|-------------|-----------------------|---------|--|
| | | | | Lower Limit | Upper Limit | Z-value | P-value | |
| Earth science | 4 | 1.08 | .16 | .12 | .21 | 7.27 | < .001 | Q(3)= 147.58, <i>p</i> < .001, <i>I</i> ² = 97.97 |
| | 8 | 1.10 | .15 | .11 | .19 | 7.46 | < .001 | Q(3)= 57.84, <i>p</i> < .001, <i>I</i> ² = 93.08 |
| | 12 | 1.15 | .21 | .17 | .26 | 9.04 | < .001 | Q(3)= 63.84, <i>p</i> < .001, <i>I</i> ² = 95.30 |
| | Overall | | .17 | .13 | .21 | 8.54 | < .001 | between groups, Q(2) = 4.16, <i>p</i> = .125 |
| Physical science | 4 | 1.11 | .05 | .02 | .09 | 3.10 | .003 | Q(3)= 35.15, <i>p</i> < .001, <i>I</i> ² = 91.47 |
| | 8 | 1.13 | .17 | .14 | .20 | 10.94 | < .001 | Q(4)= 103.27, <i>p</i> < .001, <i>I</i> ² = 96.13 |
| | 12 | 1.14 | .18 | .14 | .22 | 9.64 | < .001 | Q(3)= 20.56, <i>p</i> < .001, <i>I</i> ² = 85.41 |
| | Overall | | .13 | .06 | .21 | 3.31 | < .001 | between groups, Q(2)= 32.33, <i>p</i> < .001 |
| Life science | 4 | 1.06 | .01 | -.05 | .06 | 0.22 | .826 | Q(3)= 367.40, <i>p</i> < .001, <i>I</i> ² = 99.18 |
| | 8 | 1.10 | .04 | -.01 | .09 | 1.61 | .107 | Q(4)= 37.26, <i>p</i> < .001, <i>I</i> ² = 89.27 |
| | 12 | 1.10 | .01 | -.04 | .07 | 0.46 | .645 | Q(3)= 18.37, <i>p</i> < .001, <i>I</i> ² = 83.67 |

| | | | | | | |
|---------|-----|------|-----|------|------|---------------------------------------|
| Overall | .02 | -.01 | .05 | 1.38 | .167 | between groups, $Q(2)=0.94, p = .624$ |
|---------|-----|------|-----|------|------|---------------------------------------|

Note: Effect sizes that exceed Hyde’s criterion for non-trivial differences ($d \geq .10$) are highlighted in bold.

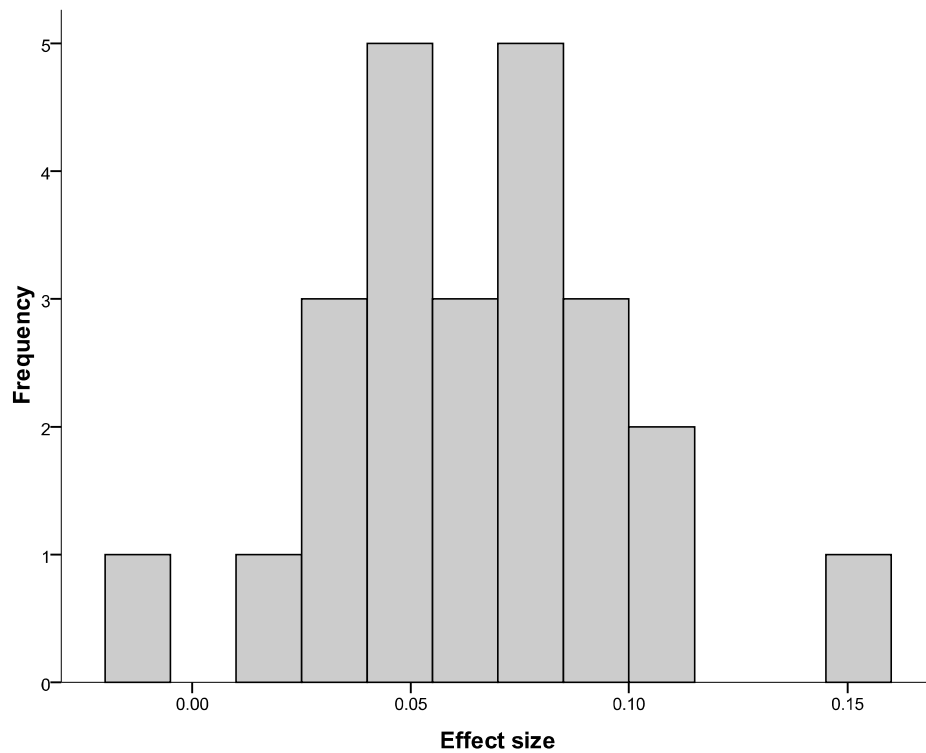


Figure 1. Histogram of observed effect sizes in NAEP mathematics assessments (1990-2011).

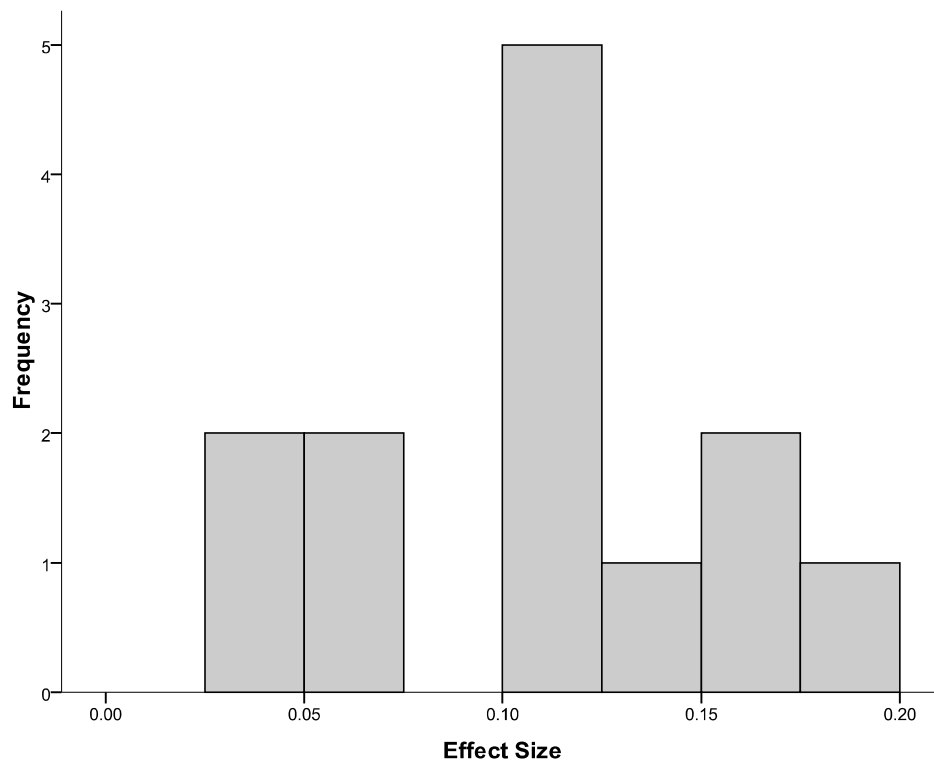


Figure 2. Histogram of observed effect sizes in NAEP science assessments (1996-2011).

Appendix

Table A1.

Descriptive Statistics, Effect Sizes and Variance Ratios for NAEP Mathematics.

| Year | Grade | Male | | Female | | Sample size | Variance ratio | Cohen's <i>d</i> |
|------|-------|-----------|----------|-----------|----------|-------------|----------------|------------------|
| | | Mean | (SD) | Mean | (SD) | | | |
| 2011 | 4 | 241.41825 | 29.76624 | 239.92438 | 28.11948 | 209,000 | 1.12 | 0.05 |
| 2009 | 4 | 240.61765 | 29.50272 | 238.69442 | 27.86402 | 168,800 | 1.12 | 0.07 |
| 2007 | 4 | 240.79044 | 29.43191 | 238.62343 | 27.74161 | 197,700 | 1.13 | 0.08 |
| 2005 | 4 | 239.11030 | 28.92446 | 236.59788 | 27.81468 | 172,000 | 1.08 | 0.09 |
| 2003 | 4 | 236.37463 | 29.06977 | 233.41351 | 27.58066 | 190,000 | 1.11 | 0.10 |
| 2000 | 4 | 226.82131 | 32.34153 | 224.30827 | 30.05055 | 13,800 | 1.16 | 0.08 |
| 1996 | 4 | 223.73966 | 31.70661 | 223.27141 | 29.95813 | 6,600 | 1.12 | 0.02 |
| 1992 | 4 | 220.89259 | 32.52064 | 218.52010 | 30.80918 | 8,700 | 1.11 | 0.07 |
| 1990 | 4 | 213.54463 | 32.73525 | 212.54085 | 30.70411 | 8,900 | 1.14 | 0.03 |
| 2011 | 8 | 284.45084 | 37.21046 | 283.23397 | 35.12125 | 175,200 | 1.12 | 0.03 |
| 2009 | 8 | 283.94915 | 37.22430 | 281.85728 | 35.48711 | 161,700 | 1.10 | 0.06 |
| 2007 | 8 | 282.40116 | 37.40132 | 280.27550 | 34.62987 | 153,000 | 1.17 | 0.06 |
| 2005 | 8 | 279.61146 | 37.14541 | 278.01277 | 35.43316 | 162,000 | 1.10 | 0.04 |
| 2003 | 8 | 278.48139 | 37.18516 | 276.63517 | 35.21715 | 153,000 | 1.11 | 0.05 |
| 2000 | 8 | 273.91265 | 39.25296 | 272.27437 | 36.79607 | 15,800 | 1.14 | 0.04 |
| 1996 | 8 | 271.43222 | 38.25208 | 269.44691 | 36.62322 | 7,100 | 1.09 | 0.05 |
| 1992 | 8 | 268.09776 | 36.78734 | 268.70292 | 35.68133 | 9,400 | 1.06 | -0.02 |
| 1990 | 8 | 263.20994 | 37.23174 | 261.87034 | 34.70190 | 8,900 | 1.15 | 0.04 |
| 2009 | 12 | 154.94494 | 34.89788 | 151.66908 | 32.47539 | 51,700 | 1.15 | 0.10 |
| 2005 | 12 | 151.31353 | 35.54736 | 148.78616 | 32.35334 | 15,100 | 1.21 | 0.07 |
| 2000 | 12 | 301.90598 | 37.44853 | 298.52331 | 33.72126 | 13,800 | 1.23 | 0.09 |
| 1996 | 12 | 302.94416 | 34.98625 | 300.34237 | 32.67684 | 6,900 | 1.15 | 0.08 |
| 1992 | 12 | 301.33159 | 34.71171 | 297.75355 | 33.04985 | 8,500 | 1.10 | 0.11 |
| 1990 | 12 | 297.08056 | 36.39719 | 291.48571 | 34.89335 | 8,900 | 1.09 | 0.16 |

Note: Effect sizes that are statistically significant at $p < .05$ are highlighted in bold.

Variance ratios (VRs) above 1.00 indicate greater male variability; VRs below 1.00 reflect greater female variability

Table A2.
Percentage of Male and Female Students Attaining the Advanced Proficiency Level for Mathematics

| Grade | Year | Male at Advanced or higher | Female at Advanced or higher | Risk ratio |
|-----------------------|------|-------------------------------|---------------------------------|-------------|
| 4 | 2011 | 7.576799 | 5.717962 | 1.33 |
| | 2009 | 6.914833 | 4.945088 | 1.40 |
| | 2007 | 6.625340 | 4.485612 | 1.48 |
| | 2005 | 5.831723 | 4.180971 | 1.39 |
| | 2003 | 4.891417 | 2.916470 | 1.68 |
| | 2000 | 3.436696 | 1.760290 | 1.95 |
| | 1996 | 3.054985 | 1.426135 | 2.14 |
| | 1992 | 2.111726 | 1.334901 | 1.58 |
| | 1990 | 1.685714 | 0.625990 | 2.69 |
| Grade 4 Ratio | | | | 1.50 |
| 8 | 2011 | 9.216519 | 7.266763 | 1.27 |
| | 2009 | 8.801046 | 7.020203 | 1.25 |
| | 2007 | 8.103602 | 5.876708 | 1.38 |
| | 2005 | 6.731102 | 5.331733 | 1.26 |
| | 2003 | 6.118808 | 4.649840 | 1.32 |
| | 2000 | 5.917239 | 4.102844 | 1.44 |
| | 1996 | 4.296899 | 3.341992 | 1.29 |
| | 1992 | 3.172883 | 2.990450 | 1.06 |
| | 1990 | 2.366303 | 1.571801 | 1.51 |
| Grade 8 Ratio | | | | 1.30 |
| 12 | 2009 | 3.520143 | 1.808225 | 1.95 |
| | 2005 | 3.062339 | 1.372165 | 2.23 |
| | 2000 | 3.233790 | 1.363867 | 2.37 |
| | 1996 | 2.538683 | 1.378109 | 1.84 |
| | 1992 | 2.082941 | 1.065519 | 1.95 |
| | 1990 | 2.287939 | 0.685097 | 3.34 |
| Grade 12 Ratio | | | | 2.13 |

Table A3.

Descriptive Statistics, Effect Sizes and Variance Ratios for NAEP Science.

| Year | Grade | Male | | Female | | Sample size | Variance ratio | Cohen's <i>d</i> |
|------|-------|-----------|----------|-----------|----------|-------------|----------------|------------------|
| | | Mean | (SD) | Mean | (SD) | | | |
| 2009 | 4 | 150.57607 | 35.71345 | 149.40869 | 34.21378 | 156,500 | 1.09 | 0.03 |
| 2005 | 4 | 152.52981 | 31.78684 | 148.65937 | 30.40344 | 172,500 | 1.09 | 0.12 |
| 2000 | 4 | 152.54469 | 35.28189 | 147.13105 | 33.46140 | 15,800 | 1.11 | 0.10 |
| 1996 | 4 | 150.85221 | 33.53692 | 149.13861 | 32.39515 | 7,300 | 1.07 | 0.04 |
| 2011 | 8 | 154.16130 | 35.05894 | 149.21276 | 33.25938 | 122,000 | 1.11 | 0.14 |
| 2009 | 8 | 151.98475 | 36.14538 | 147.99001 | 33.65030 | 151,100 | 1.15 | 0.11 |
| 2005 | 8 | 150.48951 | 36.21289 | 146.58535 | 34.31355 | 173,700 | 1.11 | 0.11 |
| 2000 | 8 | 154.35722 | 36.53652 | 147.34892 | 35.03792 | 15,800 | 1.09 | 0.17 |
| 1996 | 8 | 150.84548 | 34.83000 | 149.12814 | 32.90619 | 7,800 | 1.12 | 0.06 |
| 2009 | 12 | 152.87615 | 35.79328 | 147.15407 | 33.95205 | 11,100 | 1.11 | 0.16 |
| 2005 | 12 | 149.01019 | 34.95658 | 145.14886 | 32.58746 | 22,000 | 1.15 | 0.11 |
| 2000 | 12 | 147.66712 | 35.31678 | 145.17791 | 32.66081 | 15,800 | 1.17 | 0.07 |
| 1996 | 12 | 153.73568 | 34.34129 | 147.22612 | 32.31225 | 7,500 | 1.13 | 0.20 |

Note: Effect sizes that are statistically significant at $p < .05$ are highlighted in bold.

Variance ratios (VRs) above 1.00 indicate greater male variability; VRs below 1.00 reflect greater female variability

Table A4.

Descriptive Statistics, Effect Sizes for Males and Females Across Field of Science

| Grade | Year | Sub-domain | Male | | Female | | Variance | Effect size |
|-------|------|------------|-----------|----------|-----------|----------|------------|------------------|
| | | | Mean | (SD) | Mean | (SD) | Ratio (VR) | Cohen's <i>d</i> |
| 4 | 2009 | Earth | 151.98697 | 35.48140 | 147.94554 | 34.36776 | 1.07 | 0.12 |
| | | Physical | 150.68020 | 35.76229 | 149.30413 | 34.14173 | 1.10 | 0.04 |
| | | Life | 149.02011 | 35.45167 | 151.01676 | 34.48297 | 1.06 | -0.06 |
| | 2005 | Earth | 154.62696 | 34.23227 | 147.90774 | 32.83755 | 1.09 | 0.20 |
| | | Physical | 152.88549 | 33.46338 | 150.37863 | 31.59997 | 1.12 | 0.08 |
| | | Life | 150.07751 | 32.25717 | 147.69223 | 31.53551 | 1.05 | 0.07 |
| | 2000 | Earth | 155.29923 | 37.11476 | 146.93308 | 35.06581 | 1.12 | 0.23 |
| | | Physical | 151.73910 | 36.73263 | 147.29376 | 35.00875 | 1.10 | 0.12 |
| | | Life | 150.59632 | 36.19456 | 147.16675 | 34.57285 | 1.10 | 0.10 |
| | 1996 | Earth | 152.68547 | 35.27448 | 147.28560 | 34.50327 | 1.05 | 0.15 |
| | | Physical | 150.40359 | 35.89103 | 149.59256 | 34.06300 | 1.11 | 0.02 |
| | | Life | 149.46810 | 35.33133 | 150.53822 | 34.64733 | 1.04 | -0.03 |
| 8 | 2011 | Earth | 153.92140 | 34.84762 | 147.80144 | 33.75392 | 1.07 | 0.18 |
| | | Physical | 155.57074 | 35.11631 | 147.88813 | 32.92987 | 1.14 | 0.23 |
| | | Life | 153.30020 | 35.25709 | 151.58303 | 33.69989 | 1.09 | 0.05 |
| | 2009 | Earth | 152.87875 | 35.59819 | 147.08004 | 34.11659 | 1.09 | 0.17 |
| | | Physical | 152.65646 | 36.30124 | 147.30709 | 33.38742 | 1.18 | 0.15 |
| | | Life | 150.69328 | 35.95506 | 149.30199 | 33.95972 | 1.12 | 0.04 |
| | 2005 | Earth | 152.17593 | 36.88717 | 147.55658 | 34.80591 | 1.12 | 0.13 |
| | | Physical | 149.21860 | 39.26031 | 142.21619 | 37.12914 | 1.12 | 0.18 |
| | | Life | 150.17797 | 36.55163 | 149.13386 | 35.07240 | 1.09 | 0.03 |
| | 2000 | Earth | 155.05456 | 37.39703 | 148.12953 | 35.43240 | 1.11 | 0.19 |
| | | Physical | 155.15257 | 38.90284 | 144.44899 | 37.61012 | 1.07 | 0.28 |
| | | Life | 153.23782 | 36.96043 | 148.93848 | 35.64924 | 1.07 | 0.12 |
| | 1996 | Earth | 151.69663 | 35.79566 | 148.25032 | 34.06708 | 1.10 | 0.10 |
| | | Physical | 151.98523 | 35.94419 | 147.95250 | 33.87181 | 1.13 | 0.12 |
| | | Life | 149.35236 | 35.93675 | 150.66815 | 33.98972 | 1.12 | -0.04 |
| 12 | 2009 | Earth | 155.03106 | 35.30019 | 145.02064 | 33.96834 | 1.08 | 0.29 |
| | | Physical | 153.45769 | 36.05441 | 146.57822 | 33.56864 | 1.15 | 0.20 |
| | | Life | 150.91495 | 35.23967 | 149.09501 | 34.73195 | 1.03 | 0.05 |
| | 2005 | Earth | 147.90490 | 35.52127 | 142.57167 | 33.07296 | 1.15 | 0.16 |
| | | Physical | 151.21768 | 37.09497 | 144.72360 | 34.38433 | 1.16 | 0.18 |
| | | Life | 147.90856 | 35.57997 | 148.15184 | 33.75475 | 1.11 | -0.01 |
| | 2000 | Earth | 146.87288 | 36.40061 | 142.34830 | 33.23612 | 1.20 | 0.13 |
| | | Physical | 149.10605 | 37.42376 | 144.84199 | 35.29647 | 1.12 | 0.12 |
| | | Life | 147.02286 | 35.56595 | 148.34385 | 32.84964 | 1.17 | -0.04 |
| | 1996 | Earth | 155.85416 | 35.50721 | 146.15392 | 33.06037 | 1.15 | 0.28 |
| | | Physical | 154.14569 | 36.15524 | 146.04818 | 34.04841 | 1.13 | 0.23 |
| | | Life | 151.20767 | 34.76459 | 149.47683 | 33.20536 | 1.10 | 0.05 |

Note: Effect sizes that are statistically significant at $p < .05$ are highlighted in bold.

Table A5.

Percentage of Male and Female Students Attaining the Advanced Proficiency Level for Science

| Grade | Year | Male | Female | Gender ratio |
|-------|------|-----------------------|-----------------------|--------------|
| | | at Advanced or higher | at Advanced or higher | |

| | | | | |
|-----------------------|------|----------|----------|-------------|
| 4 | 2009 | 0.682025 | 0.512587 | 1.33 |
| | 2005 | 3.202503 | 1.853222 | 1.73 |
| | 2000 | 4.792086 | 2.544751 | 1.88 |
| | 1996 | 3.418698 | 2.695513 | 1.27 |
| Grade 4 Ratio | | | | 1.56 |
| 8 | 2011 | 2.203641 | 0.989545 | 2.23 |
| | 2009 | 2.039891 | 0.967621 | 2.11 |
| | 2005 | 4.038574 | 2.349279 | 1.72 |
| | 2000 | 5.152212 | 2.799564 | 1.84 |
| | 1996 | 3.553214 | 2.530816 | 1.40 |
| Grade 8 Ratio | | | | 1.88 |
| 12 | 2009 | 1.998120 | 0.827877 | 2.41 |
| | 2005 | 2.651007 | 1.266683 | 2.09 |
| | 2000 | 2.760126 | 1.390137 | 1.99 |
| | 1996 | 3.993961 | 1.346876 | 2.97 |
| Grade 12 Ratio | | | | 2.28 |