Psychological Bulletin
2010, Vol. 136, No. 1, 103–127

0033-2909/10/$12.00 DOI: 10.1037/a0018053

Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis

Nicole M. Else-Quest
Villanova University

Janet Shibley Hyde
University of Wisconsin—Madison

Marcia C. Linn
University of California, Berkeley

A gender gap in mathematics achievement persists in some nations but not in others. In light of the underrepresentation of women in careers in science, technology, mathematics, and engineering, increasing research attention is being devoted to understanding gender differences in mathematics achievement, attitudes, and affect. The gender stratification hypothesis maintains that such gender differences are closely related to cultural variations in opportunity structures for girls and women. We meta-analyzed 2 major international data sets, the 2003 Trends in International Mathematics and Science Study and the Programme for International Student Assessment, representing 493,495 students 14–16 years of age, to estimate the magnitude of gender differences in mathematics achievement, attitudes, and affect across 69 nations throughout the world. Consistent with the gender similarities hypothesis, all of the mean effect sizes in mathematics achievement were very small ($d < 0.15$); however, national effect sizes showed considerable variability (ds $= -0.42$ to 0.40). Despite gender similarities in achievement, boys reported more positive math attitudes and affect (ds $= 0.10$ to 0.33); national effect sizes ranged from $d = -0.61$ to 0.89. In contrast to those of previous tests of the gender stratification hypothesis, our results point to specific domains of gender equity responsible for gender gaps in math. Gender equity in school enrollment, women's share of research jobs, and women's parliamentary representation were the most powerful predictors of cross-national variability in gender gaps in math. Results are situated within the context of existing research demonstrating apparently paradoxical effects of societal gender equity and highlight the significance of increasing girls' and women's agency cross-nationally.

Keywords: gender differences, mathematics, gender equity, international data sets

The question of gender differences in mathematics achievement, attitudes, and affect is a continuing concern as scientists seek to address the underrepresentation of women at the highest levels of science, technology, mathematics, and engineering (STEM; Halpern et al., 2007; National Academy of Sciences, 2006). Stereotypes that girls and women lack mathematical ability persist, despite mounting evidence of gender similarities in math achievement (Hedges & Nowell, 1995; Hyde, Fennema, & Lamon, 1990; Hyde, Lindberg, Linn, Ellis, & Williams, 2008). Because much of the research on gender differences in math has been based on North American samples, the current study aimed to examine the magnitude of gender differences in mathematics achievement, attitudes, and affect cross-nationally. Furthermore, some have proposed the *gender stratification hypothesis*, arguing that cross-national patterns of gender differences in math achievement reflect gender inequities in educational and economic opportunities available in a given culture (Baker & Jones, 1993; Guiso, Monte, Sapienza, & Zingales, 2008; Riegle-Crumb, 2005). In the current study, we meta-analyzed two large international data sets to examine cross-national patterns of gender differences in mathematics achievement, attitudes, and affect and assessed the links of these patterns to gender equity at the national level.

Gender Differences and Similarities in Mathematics

Stereotypes about female inferiority in mathematics (Bhana, 2005; Fennema, Peterson, Carpenter, & Lubinski, 1990; Fennema & Sherman, 1977; Hyde, Fennema, Ryan, Frost, & Hopp, 1990; Li, 1999) stand in distinct contrast to the actual scientific data reported in previous studies. This discrepancy is particularly problematic because such negative stereotypes can impair math test performance and cause anxiety via stereotype threat (Blascovich, Spencer, Quinn, & Steele, 2001; Spencer, Steele, & Quinn, 1999). Reviewing evidence from research with infants and preschoolers,

Nicole M. Else-Quest, Department of Psychology, Villanova University; Janet Shibley Hyde, Department of Psychology, University of Wisconsin—Madison; Marcia C. Linn, Graduate School of Education, University of California, Berkeley.

Portions of these data were presented at the Gender Development Research Conference, San Francisco, California, April 2008, as well as at the meeting of the Society for Research in Child Development, Denver, Colorado, April 2009. This research was funded through National Science Foundation Grant REC 0635444. Any findings, conclusions, or recommendations expressed in this material are our own and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to Nicole M. Else-Quest, Villanova University, Department of Psychology, 800 Lancaster Avenue, Villanova, PA 19085. E-mail: nicole.else.quest@villanova.edu

Spelke (2005) concluded that gender similarities are the rule in the development of early number concepts. Girls earn better grades in mathematics courses through the end of high school (Dwyer & Johnson, 1997; Kenney-Benson, Pomerantz, Ryan, & Patrick, 2006; Kimball, 1989).

In the United States, gender differences in mathematics performance are declining. A meta-analysis in 1990 (Hyde, Fennema, & Lamon., 1990) found an effect size of $d = -0.05$ for the gender difference in math performance among the general population, indicating a negligible female advantage (note that positive values of d represent higher scores for males than females, whereas negative values represent higher scores for females). At that time the gender gap increased during high school. Another meta-analysis used data sets representing large probability samples of American adolescents and found $d = 0.03$ to 0.26 across the different data sets (Hedges & Nowell, 1995). More recent data indicate that the gender difference in math achievement has been eliminated. A study of statewide mathematics tests administered between 2005 and 2007 for Grades 2–11 found $d = 0.0065$, without the increased gender gap in adolescence found with earlier data (Hyde et al., 2008). These findings, for U.S. samples, are consistent with the *gender similarities hypothesis*, which maintains that males and females are similar on most, but not all, psychological variables (Hyde, 2005).

For the United States, meta-analytic studies of gender differences in attitudes and affect toward mathematics demonstrate that males tend to hold more positive attitudes about math, though the gap is small (Hyde, Fennema, Ryan, et al., 1990). Hyde et al. (1990) found that, developmentally, the gap widens during high school, when males report greater self-confidence ($d = 0.25$). Gender differences in math anxiety and self-concept have received considerable research attention, with girls tending to report higher anxiety and lower self-concept about their math abilities (Casey, Nuttall, & Pezaris, 1997; Fredricks & Eccles, 2002; Hyde, Fennema, Ryan, et al., 1990; McGraw, Lubienski, & Strutchens, 2006; Meece, Wigfield, & Eccles, 1990; Pajares & Miller, 1994); yet, these effects tend to be small to medium in magnitude. Although cross-cultural research has demonstrated similar findings (Stetsenko, Little, Gordeeva, Grasshof, & Oettingen, 2000), most of these reports have been based on North American samples; thus, it is unclear if these patterns of gender differences are generalizable to other cultures. Therefore, a focus in this paper is to estimate the magnitude of gender differences in math achievement, attitudes, and affect across two international data sets totaling 69 nations.

Others have focused not on mean gender differences but on gender differences in the upper tail of the distribution and the *greater male variability hypothesis* (Hyde & Mertz, 2009). The argument is that greater variance in test scores is displayed by males than females, so that, even if there is no average gender difference, there will still be more males among the very top performers. One statistic used to test this hypothesis is the variance ratio, VR, or the male variance divided by the female variance. Analysis of variance ratios in cross-national data (using the 2003 cycle of the Programme for International Student Assessment, or PISA) indicates that males are sometimes more variable, although the variance ratios do not indicate widely divergent differences (Hyde & Mertz, 2009; Machin & Pekkarinen, 2008). For example, VR equals 1.19 for the United States and 1.06 for the United Kingdom. However, other countries display variance ratios that do not differ significantly from 1.0 or are even significantly less than 1.0. For example, VR equals 0.99 for Denmark and 0.95 for Indonesia. Thus, despite some claims (Machin & Pekkarinen, 2008), the phenomenon of greater male variance in mathematics performance is not universal (Penner, 2008). Another method of testing the greater male variability hypothesis is to examine mean gender differences in achievement on assessments of varying difficulty level. That is, if males are overrepresented in the upper tails of the distribution, gender differences in achievement should be larger on difficult or complex problems than on easy or moderate problems. Although Hyde, Fennema, and Lamon's (1990) meta-analysis showed that males outperformed females in complex problem solving by $d = 0.29$ in high school, recent data suggest that this gap has closed. In their analyses of gender differences on the math portion of the National Assessment of Educational Progress, Hyde et al. (2008) found that gender differences were not larger on the most challenging problems. This finding provided little support for the argument that males outperform females in complex problem solving.

Assessing Mathematics Achievement, Attitudes, and Affect Cross-Nationally

Efforts to measure the mathematics achievement, attitudes, and affect of students cross-nationally have produced two large-scale recurring assessments, the Trends in International Mathematics and Science Study (TIMSS) and PISA. TIMSS is an international assessment of mathematics and science learning in eighth graders, conducted on a 4-year cycle by the International Association for the Evaluation of International Achievement (IEA), in collaboration with Statistics Canada and the Educational Testing Service. PISA is an international assessment of mathematics, reading, science, and problem-solving literacy in 15-year-olds, conducted on a 3-year cycle by the Organisation for Economic Co-operation and Development (OECD). The current study uses data from the 2003 round of TIMSS (Mullis, Martin, Gonzalez, & Chrostowski, 2004) and PISA (OECD, 2004). With regard to issues such as measurement and sampling, in light of the complex methodological issues involved in analyzing cross-national surveys of student achievement, experts regard the TIMSS and PISA data sets as high-quality (e.g., Porter & Gamoran, 2002). Nonetheless, it is important to note the differences between these two oft-cited data sets, as well as to acknowledge any limitations.

At the outset, it is important to recognize that TIMSS and PISA have explicitly different goals. TIMSS focuses on assessing the attained curriculum, or what students have learned in the classroom, as well as teacher- and school-level variables. In contrast, OECD emphasizes that the PISA test of mathematics assesses *mathematics literacy*, which is defined as "the capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen" (OECD, 2004, p. 26). In light of these differing aims, one can expect TIMSS to be more curriculum-based and PISA to be more applied. The implication that this distinction has for comparing results from the data sets is that PISA may be a more challenging assessment requiring a deeper understanding of mathematics.

The quality of any academic achievement assessment is partly determined by the depth of knowledge (DoK) or degree of difficulty assessed with its items. Ideally, a standardized indicator of DoK would be applied to TIMSS and PISA to determine whether one assessment is more challenging than the other. Because the IEA and OECD do not release to the public the full batteries of TIMSS and PISA items, respectively, it is not possible to code the DoK assessed by the test items on both assessments. However, both the IEA and OECD attempt to address this concern with their own respective coding scheme. The IEA classifies TIMSS items into three "cognitive domains," which reflect the complexity of cognitive processes required for those items (International Association for the Evaluation of International Achievement [IEA], 2007). Items in the Knowing Facts, Procedures, and Concepts domain make up 33.5% of the TIMSS items; they require recall of facts, procedures, and mathematical concepts; computation; recognition and identification of mathematical equivalence; and use of mathematics and measuring instruments. For example, "If $x = -3$, what is the value of $-3x$?" The Applying Knowledge and Understanding domain entails the application of mathematical knowledge of facts, skills, procedures, and concepts to create representations and solve routine problems. For example, "Jack wants to find how far an airplane will travel in 3.5 hours at its top speed of 965 kph. He uses his calculator to multiply 3.5 by 965 and tells his friend Jenny that the answer is 33,775 km. Jenny says 'that can't be right.' How does she know?" This domain represents the largest component of the assessment and includes 43.9% of the items. At the deepest level is the Reasoning domain, which constitutes 18.6% of the items. It requires logical, systematic thinking, such as the ability to hypothesize, analyze, evaluate, generalize, synthesize, and prove, as well as nonroutine problem solving. For example, "Twin primes are prime numbers with one other number between them. Thus, 5 and 7, 11 and 13, and 17 and 19 are pairs of twin primes. Make a conjecture about the numbers between twin primes." Based on the breakdown of items within each cognitive domain, an emphasis in the TIMSS on basic knowledge and routine problem-solving is evident.

OECD attempts to address the DoK issue by classifying PISA math items into three "competency clusters," which are based on the cognitive processes used to complete the items (OECD, 2003). Items in the Reproduction cluster make up 31% of the assessment and involve recall of facts, recognition of equivalents, manipulation of expressions, routine procedures, computations, and application of standard algorithms and skills. For example, "Write 69% as a fraction." The Connections cluster represents nearly half of the PISA math test (47%) and involves solving nonroutine problems in familiar or quasifamiliar situations. For example, "A pizzeria serves two round pizzas of the same thickness in different sizes. The smaller one has a diameter of 30 cm and costs 30 zeds. The larger one has a diameter of 40 cm and costs 40 zeds. Which pizza is better value for money? Show your reasoning." The third cluster, Reflection, comprises 22% of the items and involves advanced reasoning, insight, and creativity, requiring students to plan and implement problem-solving strategies. For example, students are shown a graph modeling combined fish growth (in kilograms) over time (in years); the item states "Suppose a fisherman plans to wait a number of years and then start catching fish from the waterway. How many years should the fisherman wait if he or she wishes to maximize the number of fish he or she can catch annually from that year on? Provide an argument to support your answer." An emphasis in the PISA on more challenging and complex mathematics is evident from the breakdown of items within each cognitive cluster. Thus, it may be surmised that PISA is a more challenging mathematics assessment than is TIMSS. Previous reports have suggested that gender differences in mathematics appear only at the level of complex problem solving (Hyde, Fennema, & Lamon, 1990), and the greater male variability hypothesis predicts larger gender differences in more challenging mathematics assessments. Thus, this apparent difference in DoK accessed on the two assessments might foreshadow larger gender differences on PISA than on TIMSS, as would be consistent with the greater male variability hypothesis.

In terms of samples, PISA differs from TIMSS insofar as it is based on age (15 years 3 months to 16 years 2 months) rather than grade level. Because the majority of nations begin formal schooling at age 6, most of the eighth graders assessed in TIMSS were between 14 and 15 years of age, approximately one year younger than those in the PISA sample. The developmental importance of this age difference in samples is unclear. Although previous reports have demonstrated a widening of the gender gap in math achievement and attitudes during this developmental period (Hyde, Fennema, & Lamon, 1990; Hyde, Fennema, Ryan, et al., 1990), recent data suggest that the gap has closed (Hyde et al., 2008). Regardless, students of this age range (14–16 years) are old enough to be capable of complex mathematical problem solving.

Both TIMSS and PISA hold participating nations to strict standards in terms of sampling and test administration. This provides confidence in the reliability and quality the data. Yet, these standards incur major costs on the nations that participate in these voluntary assessments (Hutchison & Schagen, 2007); thus, nations must have organized formal schooling systems and enjoy a certain level of prosperity in order to participate. None of the nations in the TIMSS or PISA data sets are characterized as low in human development (i.e., having a Human Development Index, or HDI, of less than 0.50), according to the1995 *Human Development Report* (UN Development Programme, 2003). Instead, over half of the nations included in TIMSS and 84% of the nations in PISA sample were characterized as high in human development in 2003. It is noteworthy that TIMSS represents a more diverse and less developed sample of nations than does PISA, as the PISA sample reflects membership in the OECD. The overrepresentation of developed nations in the TIMSS and PISA samples is a limitation that is difficult to overcome with international academic achievement assessments.

In sum, TIMSS and PISA differ somewhat with regard to the sample age and level of development and the difficulty level of the test items. These differences could be reflected in the findings of the current study, insofar as results from TIMSS and PISA may vary. Insofar as the greater male variability hypothesis maintains that males should be overrepresented at the upper tails of the distribution in mathematics ability, a male advantage in the most difficult math problems should exist. Thus, if the greater male variability hypothesis is valid, we would expect larger gender differences in more challenging assessments of math achievement, such as PISA, or on problems assessing deeper levels of processing, such as the TIMSS cognitive domain of Reasoning.

Theoretical Framework

What factors might contribute to gender differences or similarities in math achievement, attitudes, and affect? Dozens of explanations have been proposed, including hormones and prenatal brain differentiation, stereotype threat, and other factors (Byrnes, 2005; Ceci & Williams, 2007; Halpern et al., 2007). Of primary interest here is a sociological hypothesis proposed by Baker and Jones (1993), who argued that girls' poorer math achievement and more negative math attitudes are the result of societal gender stratification. The gender stratification hypothesis proposes that in patriarchal cultures, male students link their achievement to future opportunities and outcomes. As a result of the decreased opportunities afforded to females, girls do not perceive such a link and thus do not achieve as boys do in domains that they perceive to be less useful. Baker and Jones (1993) argued that

> female students, who are faced with less opportunity, may see mathematics as less important for their future and are told so in a number of ways by teachers, parents, and friends. In short, opportunity structures can shape numerous socialization processes that shape performance. (p. 92)

Broadly, the gender stratification hypothesis proposes that, where there is more societal stratification based on gender, and thus more inequality of opportunity, girls will report less positive attitudes and more negative affect and will perform less well on mathematics achievement tests than will their male peers. Yet, where there is greater gender equity, gender similarities in math will be evident.

Three theoretical approaches within psychology provide some insight into socialization processes that might account for the effects proposed by the gender stratification hypothesis. Eccles and her colleagues (e.g., Eccles, 1994; Jacobs, Davis-Kean, Bleeker, Eccles, & Malanchuk, 2005; Meece, Eccles-Parsons, Kaczala, Goff, & Futterman, 1982) have proposed and tested an expectancy-value theoretical model to explain the gender gap in mathematics achievement, attitudes, and affect and the underrepresentation of women in careers in science and engineering. According to the Eccles model, people do not undertake a challenge unless they value it and have some expectation of success. Perceptions of the value of the task (e.g., taking a challenging mathematics course) are shaped by the cultural milieu (e.g., gender segregation of occupations, cultural stereotypes about the subject matter) and the person's short-term and long-term goals (e.g., becoming an elementary school teacher and thinking one does not need advanced mathematics or becoming a civil engineer and knowing that one does). Expectations of success are shaped by the person's aptitude, relevant past events such as grades in the subject and scores on standardized tests, the person's interpretations of and attributions for these events, and the person's self-concept of ability. Sociocultural forces such as parents' and teachers' attitudes and expectations, including stereotypes, also shape self-concept and attitudes toward the subject; empirical research on the awareness of negative stereotypes supports this link (Aronson & McGlone, 2008; Pinel, 1999). According to the expectancy-value model, if a girl believes that the career opportunities available to or appropriate for women do not require mathematics skills, she is less likely to invest in developing her mathematics skills by working hard in her required math courses or by taking elective math courses. She may see math as less useful or valuable and may think she is not capable of doing math. The theory has received abundant empirical support (e.g., Eccles, 1994; Frome & Eccles, 1998) and provides a clear model for why cultural inequities in educational or career opportunities have an adverse impact on girls and women considering STEM careers. Eccles's (1994) expectancy-value theoretical model is consistent with the gender stratification hypothesis in maintaining that individuals do not engage in tasks that are perceived to have little value and arguing that individuals make cost–benefit judgments regarding their academic choices.

Another psychological theory that is consistent with the gender stratification hypothesis is cognitive social learning theory (Bandura, 1986; Bussey & Bandura, 1999), which maintains that a number of social processes contribute to the development of gender-typed behavior, including reinforcements, modeling, and cognitive processes, such as self-efficacy. Role models and socializing agents, as well as perceptions of gender-appropriate behavior, have an important influence on an individual's academic choices. As with Eccles's model, this theory also emphasizes the role of self-efficacy in gender-typed behaviors, such as choosing to major in physics. This theory maintains that girls are attentive to the behaviors that women in their culture engage in and thus feel efficacious in and model those behaviors. That is, if girls observe that women in their culture do not become engineers or scientists, they may believe that such careers (and, by extension, STEM subjects) are outside the realm of possibilities for girls and feel anxious about and/or avoid these subjects. In emphasizing the roles of observational learning and the internalization of cultural norms, cognitive social learning theory provides an individual-level explanation of why girls and women make gendered educational and vocational choices that recapitulate societal-level gender stratification.

Social structural theory (sometimes referred to as social role theory; Eagly, 1987; Eagly & Wood, 1999) is another relevant psychological theory in that it maintains that psychological gender differences are rooted in sociocultural factors, such as the gendered division of labor. A society's gendered division of labor fosters the development of gender differences in behavior by affording different restrictions and opportunities to males and females on the basis of their social roles. Accordingly, if girls are expected to care for younger siblings or prepare meals rather than learn algebra, their access to formal schooling may be limited. Eagly and Wood's (1999) cross-cultural analyses tend to support social structural theory, demonstrating substantial correlations between composite indicators of gender equity and gender differences in mate preferences (including earning capacity, domestic skills, and age). Although Eagly and Wood did not analyze gender differences in mathematics, social structural theory can be applied to access to mathematics education. That is, if the cultural roles that women fulfill do not include math, girls may face both structural obstacles (e.g., formal access to education is limited to boys) and social obstacles (e.g., stereotypes that math is a male domain) that impede their mathematical development. According to social structural theory, across nations, gender equity in educational and employment opportunities should be associated with gender similarities in mathematics achievement, attitudes, and affect.

Previous Tests of the Gender Stratification Hypothesis

To test the gender stratification hypothesis, Baker and Jones (1993) used cross-national data from the Second International Mathematics Study (SIMS; an early precursor to TIMSS), which administered the same mathematics test to a representative sample of 13-year-old students in each of 19 countries around the globe in 1982. Using United Nations Development Programme (UNDP) data, Baker and Jones constructed several variables measuring gender equity in the countries (e.g., percentage of females in higher education and percentage of females in the labor force). They then computed correlations between the magnitude of gender differences in mathematics achievement and each of the gender inequality variables, essentially testing whether gender equity moderated the gender difference in math performance on the SIMS. Their results showed that boys significantly outperformed girls in seven of the 19 nations and that girls significantly outperformed boys in four. There were no significant gender differences in the remaining eight nations. Baker and Jones also found support for the gender stratification hypothesis; a smaller gender gap in math was significantly correlated with greater women's labor force participation, the percentage of women in higher education, and the percentage of women working in the industrial and service economic sectors. Gender occupational segregation, the ratio of women in university–nonuniversity programs, and the percentage of women working in the agricultural economic sector were not significantly correlated with the gender gap in math performance but were significantly correlated with perceived parental encouragement for math achievement. Although these findings provide evidence of cultural influence—specifically, the importance of equal opportunity in a culture—they are limited by several methodological constraints.

The SIMS is now more than 26 years old, and international math assessments have improved greatly since then in terms of items and administration (Mullis & Martin, 2007). The sample of nations participating in the SIMS was primarily limited to developed nations, in which there tends to be relatively greater gender equity (Hausmann, Tyson, & Zahidi, 2007; UN Development Programme [UNDP], 1995; but see Riegle-Crumb, 2005, for an opposing view). What do analyses with more recent data suggest about the gender stratification hypothesis?

Riegle-Crumb (2005) used cross-national data from the Third International Mathematics and Science Study (TIMSS-1995) to test the gender stratification hypothesis. An update to the SIMS and precursor to the TIMSS 2003 used in the current study, TIMSS-1995 included more than twice as many countries as did the 1982 SIMS assessment. Riegle-Crumb argued that, when girls witness a lack of women in power, the status quo in gender inequality is maintained by limiting their expectations of success and achievement. To measure gender equity in the domains of labor force and government representation, Riegle-Crumb used variables similar to those used by Baker and Jones and added measures to assess gender equity in the home and family (indicated by fertility rate and availability of legal abortion). Her analyses demonstrated that boys outperformed girls in 80% of the nations sampled. In addition, although greater female representation in national governments predicted a smaller gender difference in math achievement, women's economic development and relative status in the home and family did not significantly predict the gender gap in math achievement. These findings indicate that the picture is more complex than first theorized. They provide mixed support for the gender stratification hypothesis, suggesting that specific domains of gender equity are important in understanding how societal gender inequities are linked to the gender gap in math performance.

In a more widely publicized report, Guiso et al. (2008) tested the gender stratification hypothesis using different indicators of both math achievement and gender equity than those Baker and Jones and Riegle-Crumb had used in their studies. To measure the gender gap in math performance, Guiso et al. used the 2003 PISA data. To assess gender equity, they used the World Economic Forum's Gender Gap Index (GGI; Hausmann et al., 2007), a composite indicator that includes many of the social, economic, and political variables used by both Baker and Jones (1993) and Riegle-Crumb (2005). Guiso et al. found that girls' performance was, on average, only 2% lower than boys' performance. Their analyses supported the gender stratification hypothesis in that the GGI significantly predicted the magnitude of the (albeit small) gender gap in math performance. In addition, analyses of genetic distance suggested that biological differences across countries could not explain the cross-national pattern of gender differences in math performance. Although this study suggests that a multidimensional indicator of gender equity is a good predictor of gender differences in math test scores, it does not shed light on the specific domains of gender equity that are most relevant to math achievement and leaves the debate about the mechanisms of the gender stratification hypothesis unresolved.

Several studies have assessed the gender stratification hypothesis with regard to greater male variability, though with somewhat conflicting findings. Machin and Pekkarinen (2008), using 2003 PISA data, found no correlation between VR in math achievement and national gender equity. In contrast, Hyde and Mertz (2009) used PISA 2003 data and found a negative correlation between gender equity and the ratio of males to females scoring above the 95th percentile on the 2003 PISA. Similarly, Penner (2008) used TIMSS-1995 data and found that the proportion of girls scoring above the 95th percentile was linked to national gender equity.

Considered together, the results of previous tests of the gender stratification hypothesis indicate the need for several methodological revisions. With regard to cross-national patterns of gender differences in math achievement, it must first be demonstrated that a gap still exists. There is increasing agreement among researchers that the gender difference in math performance is very small in some nations, such as the United States (Hedges & Nowell, 1995; Hyde, Fennema, & Lamon, 1990; Hyde et al., 2008); however, it is unclear to what extent this gender gap varies across countries. Similarly, the extent of cross-national variations in gender differences in math attitudes and affect is not well understood, despite their links to math achievement (e.g., Bandura, Barbaranelli, Caprara, & Pastorelli, 2001; Eccles, 1994). Also, the sample of nations included in the analyses should be maximized, as occurs in the more recent cross-national assessments. Use of multiple cross-national assessments of math achievement would ensure a more reliable and thorough test of the gender stratification hypothesis.

Defining Gender Equity

In terms of assessing gender equity across nations, these findings point to the importance of how gender equity is defined and measured in the context of predicting gaps in math achievement, attitudes, and affect. Various composite indicators of national gender equity have been developed, though there is controversy regarding their use and relative strengths. Table 1 lists the most widely used composite indicators, their components, and limitations. It is noteworthy that one of the most widely used indices, the Gender Development Index (GDI; UNDP, 1995), is not actually an indicator of gender equity per se. The GDI, which was first published in the 1995 *Human Development Report* (*HDR 1995*) focusing on women's empowerment, has been and continues to be misconstrued as an indicator of national gender equity and misused as such in a variety of popular press and academic publications (Schüler, 2006). For example, Eagly and Wood (1999) as well as Schmitt, Realo, Voracek, and Allik (2008) used the GDI, among other indicators, as a measure of gender equality in their analyses. In fact, the GDI is an indicator of human development that is discounted for gender inequity. Also developed by the UNDP for *HDR 1995*, the Gender Empowerment Measure (GEM) assesses national gender equity in political, health, and economic domains. Although the GEM is perhaps the most widely used composite indicator, its utility is limited by its omission of gender equity in education, which is of particular relevance for the current study.

In response to the challenge of measuring societal or national gender equity, social scientists have developed several composite indices, including the Gender Equality Index (GEQ; White, 1997), the Standardized Index of Gender Equality (SIGE; Dijkstra, 2002), and the GGI used by Guiso et al. (2008). Like the GEM, these types of composite indices typically are computed using male-to-female ratios in a variety of domains, such as health (e.g., life expectancy and legal access to elective abortion), education (e.g., enrollment ratios and literacy rates), economics (e.g., earned income, economic activity rates, labor market participation), and politics (e.g., proportion of parliamentary seats held by women), with some domains being weighted more than others (Dijkstra, 2006). Because these composite indices are each computed differently, reflecting some domains more than or instead of others, their predictive validity also varies.

Some aspects of gender equity may be more germane to math achievement than others; for example, equal access to formal schooling (at all levels) surely has a profound impact on girls' math skills, but women's greater life expectancy is probably less relevant. A cross-national measure of women's involvement in STEM careers would test the gender stratification hypothesis more directly. Thus, indicators in multiple individual domains germane to math achievement, attitudes, and affect—including educational, economic, and political—as well as composite measures (e.g., GEQ, SIGE, GGI, GEM) should be used to provide the broadest test of the gender stratification hypothesis. This approach would also indicate the societal domains with the strongest links to the gender gap in math, thus providing insight into the mechanisms in question.

The Current Study

There were two major goals in the current study. The first was to use meta-analysis to estimate the magnitude of gender differences in mathematics achievement, attitudes, and affect using the most recent data from TIMSS and PISA. Because recent research

Table 1
Composite and Domain-Specific Indicators of Societal Gender Equity

Indicator type	Description
Composite	
Gender Empowerment Measure (GEM)	Includes women's and men's percentage shares of parliamentary seats; positions as legislators, senior officials, and managers; and professional and technical positions; includes women's and men's estimated earned income; omits education domain
Gender Equality Index (GEQ)	Assesses underlying gender equality in Gender Development Index (GDI); calculated as GDI/HDI; omits political domain
Standardized Index of Gender Equality (SIGE)	Includes relative female-to-male access to education, life expectancy, economic activity rate; women's share in higher labor market occupations; women's share in parliamentary seats; weights economic domain heavily
Gender Gap Index (GGI)	Composed of four subindices based on economic participation and opportunity, educational attainment, political empowerment, and health/survival
Domain-specific	
Primary enrollment ratio	% of female population of official school age enrolled in primary education/% of male population of official school age enrolled in primary education
Secondary enrollment ratio	% of female population of official school age enrolled in secondary education/% of male population of official school age enrolled in secondary education
Tertiary enrollment ratio	% of female population of official school age enrolled in tertiary education/% of male population of official school age enrolled in tertiary education
Economic activity rate ratio	the ratio of the proportion of females age 15 or older who supply or are available to supply labor for the production of goods and services to the proportion of males age 15 or older who supply or are available to supply labor for the production of goods and services
Women's share of higher labor market positions	% of higher labor market positions (technical and professional, as well as administrative and management positions) held by women
Women's share of research positions	% of research positions (according to International Labour Organization, 1990) held by women
Women's share of parliamentary seats	% of parliamentary seats held by women

Note. For all indicators, higher values indicate higher status of women. HDI = Human Development Index.

with North American samples has indicated little evidence of gender differences in math achievement, and as consistent with the gender similarities hypothesis, we predicted that the data would show a pattern of gender similarities in math achievement in many nations. However, because TIMSS and PISA appear to differ in the difficulty level of the items, it was expected that mean effect sizes from PISA would be slightly larger than those found with TIMSS. On the basis of previous findings in math attitudes and affect, we predicted that males would show more positive math attitudes and affect. We also predicted that there would be variability in the direction and magnitude of gender differences in math achievement, attitude, and affect across nations.

The second goal in the current study was to explain cross-national variability in these gender differences, using the most recent data from two international data sets (TIMSS and PISA), expanding upon the findings of Baker and Jones (1993), Riegle-Crumb (2005), and Guiso et al. (2008), and testing the gender stratification hypothesis. Our study improves upon those reports insofar as it includes data that (a) are from the most recent years available; (b) are from a larger sample of nations; (c) are from two international studies; (d) assess gender differences in math attitudes and affect in addition to achievement; and (e) reflect multiple domains of societal gender equity. This study is therefore well positioned to provide powerful clues regarding the cultural factors associated with narrowing or perhaps reversing the gender gaps in mathematics achievement, attitudes, and affect. We hypothesized that global composite indicators of societal gender equity (including GEM, GEQ, SIGE, and GGI) would explain the cross-national variation in the math gender gap. To focus on specific mechanisms, we predicted that indicators from domains most germane to mathematics achievement, attitudes, and affect would be the most robust moderators of the gender gap; these indicators include women's representation in scientific research, technical/professional, and administrative/management jobs, as well as girls' access to primary, secondary, and tertiary education. It was expected that indicators that most directly and closely reflect the mechanisms specified by theoretical models such as Eccles's expectancy-value theory, Bandura's social cognitive theory, and Eagly and Wood's social structural theory would be the best predictors of gender differences in math. For example, the indicator of women's share of research positions was expected to be the strongest predictor of the gender gap in math because it measures women's STEM representation, which signifies the STEM-related opportunity structures available to girls and women. Indicators of societal gender equity that seem less directly related (but still relevant) to girls' and women's opportunities in STEM—such as parliamentary representation—were expected to be less robust predictors of gender differences in math achievement, attitudes, and affect.

Method

International Data Sets of Mathematics Achievement, Attitudes, and Affect

TIMSS 2003. TIMSS 2003 was conducted in 49 countries, although two (Syrian Arab Republic and Yemen) were excluded because of sampling problems; in addition, the data from one nation (Argentina) were not available for the final TIMSS report (Mullis et al., 2004). The TIMSS data used in the meta-analysis are from the remaining 46 countries and represent the achievement, attitudes, and affect of 219,612 students. Countries and their sample sizes appear in Table 2.

Achievement. To assess mathematics achievement, TIMSS 2003 included five content domains in addition to a Math composite, which comprises the five content domains; these content domains are Number, Algebra, Measurement, Geometry, and Data. The Number content domain includes whole numbers, fractions, decimals, integers, ratios, proportion, and percentages. The content domain of Algebra assesses understanding of patterns, algebraic expressions, equations and formulas, and relationships. Measurement includes the topics of attributes and units and tools, techniques, and formulas. The Geometry content domain assesses knowledge of lines and angles, two- and three-dimensional shapes, congruence and similarity, locations and spatial relationships, and symmetry and transformations. The Data domain includes the topics of data collection and organization, data representation, data interpretation, and uncertainty and probability. As described in the Introduction, TIMSS items fall into three cognitive domains (Knowing, Applying, and Reasoning). Gender differences in these three cognitive domains appear in Mullis, Martin, and Foy (2005) and are meta-analyzed in the current study.

Attitudes and affect. In addition to assessing achievement, TIMSS administered two scales of students' math attitudes and affect (Martin, Mullis, & Chrostowski, 2004). The first scale, Self-Confidence in Mathematics, is based on the mean of four items with which students rate their agreement (e.g., "I learn things quickly in mathematics"). The second scale, Students' Valuing Mathematics, is composed of the mean of seven items with which students rate their agreement (e.g., "I need to do well in mathematics to get the job I want"). Effect sizes for gender differences in achievement and attitudes on the TIMSS appear in Table 2.

PISA 2003. PISA 2003 included 41 countries and represented the achievement, attitudes, and affect of 273,883 students. The current study meta-analyzed findings from the 2003 assessment because it focused predominantly on mathematics. Countries and their sample sizes are displayed in Table 2.

Achievement. The mathematics section of PISA includes four content domains in addition to a Math composite that comprises the content domains; these content domains are Quantity, Space/Shape, Change/Relationships, and Uncertainty. The Quantity content domain assesses understanding of numeric phenomena, quantitative relationships, and patterns; it is somewhat comparable to the TIMSS content domain of Number. The content domain of Space/Shape assesses understanding of spatial and geometric phenomena and relationships; it is somewhat comparable to the TIMSS content domain of Geometry. The Change/Relationships content domain assesses understanding of mathematical manifestations of change, functional relationships, and dependency among variables; it is to some extent comparable to the TIMSS content domain of Algebra. The content domain of Uncertainty assesses understanding of probabilities and statistics; this content domain is somewhat comparable to the TIMSS domain of Data.

Attitudes and affect. PISA assessed attitudes and affect about mathematics with five scales (OECD, 2003). Extrinsic Motivation is based on the mean of four items; students rated their agreement with statements such as "I will learn many things in Mathematics that will help me get a job." Intrinsic Motivation is composed of

Table 2
Sample Sizes and Unweighted Effect Sizes (d) for Gender Differences in Math Achievement, Attitudes, and Affect, by Nation

	TIMSS									
Nation	N_F	N_M	Math	Algebra	Data	Geometry	Measurement	Number	SCM	VM
Armenia	3,035	2,691	−0.12	−0.16	−0.14	−0.13	−0.01	−0.13	0.04	0.18
Australia	2,443	2,348	0.15	0.06	0.12	0.15	0.18	0.20	0.19	0.17
Austria	b	b	b	b	b	b	b	b	b	b
Bahrain	2,100	2,100	−0.42	−0.55	−0.36	−0.43	−0.13	−0.28	−0.12	0.11
Belgium	2,684	2,286	0.14	0.07	0.15	0.14	0.18	0.22	0.25	0.38
Botswana	2,627	2,524	−0.04	−0.14	−0.01	0.17	0.04	−0.12	0.10	−0.12
Brazil	b	b	b	b	b	b	b	b	b	b
Bulgaria	1,976	2,141	0.01	−0.10	0.09	0.04	0.03	−0.01	0.21	0.07
Canada	b	b	b	b	b	b	b	b	b	b
Chile	3,061	3,316	0.18	0.10	0.16	0.19	0.28	0.20	0.29	0.19
Chinese Taipei	2,582	2,797	−0.07	−0.12	−0.05	−0.13	0.03	−0.06	0.25	0.20
Cyprus	1,961	2,041	−0.19	−0.31	−0.19	−0.17	−0.09	−0.17	−0.03	0.00
Czech Republic	b	b	b	b	b	b	b	b	b	b
Denmark	b	b	b	b	b	b	b	b	b	b
Egypt	3,264	3,831	−0.01	−0.10	0.01	0.02	0.10	0.01	0.22	0.12
England	1,415	1,415	−0.01	−0.05	0.00	0.05	0.03	0.03	0.32	0.23
Estonia	2,020	2,020	−0.03	−0.02	−0.08	0.02	−0.05	−0.07	−0.06	−0.02
Finland	b	b	b	b	b	b	b	b	b	b
France	b	b	b	b	b	b	b	b	b	b
Germany	b	b	b	b	b	b	b	b	b	b
Ghana	2,295	2,805	0.19	0.12	0.13	0.33	0.11	0.13	0.20	−0.04
Greece	b	b	b	b	b	b	b	b	b	b
Hong Kong	2,486	2,486	−0.03	−0.06	−0.06	0.03	0.03	−0.03	0.43	0.19
Hungary	1,651	1,651	0.09	−0.04	0.06	0.14	0.19	0.11	0.15	0.11
Iceland	b	b	b	b	b	b	b	b	b	b
Indonesia	2,881	2,881	−0.01	−0.09	0.04	0.12	0.00	−0.07	0.11	−0.07
Iran, Islamic Republic of	1,977	2,965	−0.12	−0.38	−0.05	−0.19	0.11	−0.08	0.10	0.10
Ireland	b	b	b	b	b	b	b	b	b	b
Israel	2,245	2,073	0.09	−0.03	0.12	0.01	0.18	0.13	0.15	0.19
Italy	2,139	2,139	0.06	−0.04	0.15	0.07	0.15	0.08	0.18	0.27
Japan	2,379	2,477	0.03	−0.05	0.07	−0.04	0.00	0.07	0.38	0.23
Jordan	2,200	2,289	−0.30	−0.38	−0.26	−0.21	−0.18	−0.27	0.07	0.05
Korea, Republic of	2,548	2,761	0.07	0.02	0.14	0.09	0.05	0.08	0.34	0.23
Latvia	1,779	1,851	−0.07	−0.19	−0.16	−0.08	0.10	−0.03	0.09	0.13
Lebanon	2,174	1,640	0.15	−0.01	0.08	0.21	0.30	0.10	0.16	0.10
Liechtenstein	b	b	b	b	b	b	b	b	b	b
Lithuania	2,482	2,482	−0.05	−0.19	0.02	−0.04	0.03	−0.04	0.13	0.02
Luxembourg	b	b	b	b	b	b	b	b	b	b
Macao	b	b	b	b	b	b	b	b	b	b
Macedonia	1,908	1,985	−0.09	−0.21	−0.05	−0.08	0.02	−0.09	0.07	0.14
Malaysia	2,657	2,657	−0.09	−0.17	−0.06	0.01	−0.02	−0.14	0.02	−0.14
Mexico	b	b	b	b	b	b	b	b	b	b
Moldova, Republic of	2,057	1,976	−0.12	−0.20	−0.08	−0.10	0.00	−0.14	−0.02	−0.06
Morocco	1,472	1,472	0.18	0.03	0.25	0.23	0.22	0.25	0.26	0.05
Netherlands	1,502	1,563	0.10	−0.03	0.11	0.03	0.19	0.15	0.39	0.46
New Zealand	1,977	1,824	−0.03	−0.12	−0.10	−0.05	0.06	0.04	0.23	0.18
Norway	2,067	2,067	−0.04	−0.10	−0.04	−0.06	0.06	−0.03	0.24	0.24
Palestinian National Authority	2,946	2,411	−0.09	−0.26	−0.18	−0.08	0.13	−0.04	0.01	−0.05
Philippines	4,012	2,905	−0.15	−0.19	−0.14	0.02	−0.04	−0.20	−0.02	−0.16
Poland	b	b	b	b	b	b	b	b	b	b
Portugal	b	b	b	b	b	b	b	b	b	b
Romania	2,134	1,970	−0.04	−0.15	0.00	0.06	0.03	−0.06	0.14	0.05
Russian Federation	2,287	2,380	−0.04	−0.17	0.03	−0.05	0.06	−0.02	−0.08	0.00
Saudi Arabia	1,847	2,448	0.13	−0.04	−0.13	0.01	0.27	0.29	0.06	0.19
Scotland	1,758	1,758	−0.07	−0.11	−0.05	−0.07	0.00	−0.05	0.27	0.15
Serbia	2,105	2,191	−0.08	−0.18	0.04	−0.09	0.02	−0.06	0.02	0.04
Singapore	2,949	3,069	−0.13	−0.16	−0.04	−0.11	−0.06	−0.14	0.20	0.09
Slovak Republic	2,023	2,192	0.00	−0.13	0.16	0.09	0.08	0.00	0.12	0.17
Slovenia	1,789	1,789	−0.06	−0.21	−0.04	−0.08	0.08	−0.01	0.10	0.09
South Africa	4,566	4,386	0.02	0.02	−0.03	−0.01	0.05	0.01	0.05	−0.06
Spain	b	b	b	b	b	b	b	b	b	b
Sweden	2,171	2,085	0.00	−0.05	−0.01	−0.05	0.09	0.03	0.34	0.22
Switzerland	b	b	b	b	b	b	b	b	b	b

PISA											
N_F	N_M	Math	Quantity	Space	Change	Uncertainty	EM	IM	Anx.	MSC	MSE
a	a	a	a	a	a	a	a	a	a	a	a
6,171	6,380	0.06	0.01	0.11	0.04	0.07	0.24	0.23	−0.31	0.34	0.37
2,294	2,303	0.08	0.04	0.17	0.05	0.08	0.58	0.40	−0.36	0.44	0.46
a	a	a	a	a	a	a	a	a	a	a	a
4,210	4,586	0.07	0.01	0.16	0.07	0.07	0.32	0.20	−0.32	0.35	0.36
a	a	a	a	a	a	a	a	a	a	a	a
2,383	2,067	0.16	0.17	0.15	0.16	0.18	0.10	0.17	−0.34	0.33	0.30
a	a	a	a	a	a	a	a	a	a	a	a
13,037	12,688	0.13	0.05	0.20	0.15	0.15	0.12	0.17	−0.33	0.33	0.34
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
3,118	3,202	0.16	0.06	0.25	0.13	0.18	0.26	0.26	−0.26	0.36	0.42
2,147	2,071	0.18	0.10	0.16	0.21	0.24	0.43	0.29	−0.38	0.48	0.45
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
2,906	2,890	0.09	0.04	0.03	0.12	0.14	0.36	0.34	−0.39	0.45	0.56
2,262	2,038	0.09	0.02	0.17	0.04	0.12	0.35	0.24	−0.39	0.37	0.31
2,290	2,322	0.09	0.01	0.10	0.11	0.19	0.45	0.37	−0.38	0.50	0.46
a	a	a	a	a	a	a	a	a	a	a	a
2,392	2,234	0.21	0.23	0.19	0.17	0.23	0.26	0.31	−0.26	0.30	0.44
2,231	2,247	0.04	−0.03	0.04	0.01	0.12	0.20	0.27	−0.28	0.35	0.30
2,251	2,514	0.08	0.02	0.14	0.10	0.09	0.22	0.12	−0.20	0.24	0.35
1,620	1,730	−0.17	−0.30	−0.16	−0.10	−0.08	0.06	0.07	−0.27	0.22	0.25
5,419	5,342	0.04	0.02	0.18	0.04	−0.07	−0.05	0.08	−0.13	0.18	0.08
a	a	a	a	a	a	a	a	a	a	a	a
1,925	1,955	0.17	0.10	0.27	0.14	0.17	0.32	0.04	−0.28	0.23	0.30
a	a	a	a	a	a	a	a	a	a	a	a
6,041	5,598	0.19	0.12	0.17	0.20	0.25	0.23	0.11	−0.17	0.14	0.36
2,433	2,273	0.08	0.03	0.08	0.06	0.14	0.31	0.26	−0.26	0.36	0.31
a	a	a	a	a	a	a	a	a	a	a	a
2,206	3,238	0.25	0.24	0.23	0.25	0.24	0.20	0.16	−0.14	0.26	0.20
2,408	2,219	0.03	0.03	0.14	−0.01	0.00	0.18	0.20	−0.26	0.31	0.34
a	a	a	a	a	a	a	a	a	a	a	a
162	170	0.29	0.23	0.36	0.24	0.32	0.89	0.60	−0.61	0.77	0.65
a	a	a	a	a	a	a	a	a	a	a	a
1,992	1,931	0.19	0.09	0.28	0.14	0.23	0.42	0.32	−0.44	0.46	0.43
642	608	0.24	0.19	0.24	0.20	0.20	0.24	0.34	−0.46	0.47	0.38
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
15,546	14,437	0.13	0.13	0.18	0.08	0.06	0.03	0.16	−0.13	0.15	0.18
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
1,956	2,036	0.06	−0.04	0.09	0.06	0.11	0.50	0.34	−0.38	0.55	0.59
2,256	2,255	0.15	0.12	0.17	0.17	0.12	0.17	0.23	−0.31	0.35	0.37
2,015	2,049	0.07	0.00	0.07	0.04	0.10	0.23	0.25	−0.36	0.42	0.37
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
2,197	2,186	0.06	0.02	0.12	0.08	0.03	0.05	0.11	−0.03	0.18	0.17
2,416	2,192	0.14	0.15	0.16	0.13	0.12	0.06	0.03	−0.22	0.21	0.24
a	a	a	a	a	a	a	a	a	a	a	a
3,007	2,967	0.11	0.07	0.18	0.03	0.09	0.09	0.01	−0.16	0.07	0.33
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
2,228	2,177	0.01	−0.03	0.03	0.01	0.06	0.21	0.18	−0.04	0.14	0.27
a	a	a	a	a	a	a	a	a	a	a	a
3,585	3,761	0.20	0.13	0.30	0.16	0.20	0.23	0.17	−0.25	0.30	0.33
a	a	a	a	a	a	a	a	a	a	a	a
a	a	a	a	a	a	a	a	a	a	a	a
5,482	5,308	0.10	0.05	0.20	0.08	0.09	0.09	0.03	−0.34	0.25	0.28
2,310	2,314	0.07	0.04	0.10	0.01	0.09	0.32	0.19	−0.30	0.35	0.27
4,063	4,357	0.17	0.07	0.23	0.13	0.20	0.67	0.58	−0.44	0.67	0.54

(table continues)

Table 2 (*continued*)

Nation	TIMSS										
	N_F	N_M	Math	Algebra	Data	Geometry	Measurement	Number	SCM	VM
Thailand	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
Tunisia	2,613	2,318	0.40	0.21	0.40	0.30	0.37	0.39	0.27	0.12
Turkey	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
United Kingdom	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
United States	4,634	4,278	0.06	−0.01	0.01	0.08	0.15	0.10	0.26	0.05
Uruguay	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]

Note. Positive values of *d* represent higher scores for males than females, whereas negative values represent higher scores for females. TIMSS = Trends in International Mathematics and Science Study; PISA = Programme for International Student Assessment; SCM = Self-Confidence in Mathematics; VM = Valuing Mathematics; EM = Extrinsic Motivation; IM = Intrinsic Motivation; Anx. = Math Anxiety; MSC = Mathematics Self-Concept; MSE = Mathematics Self-Efficacy.
[a] Nation did not participate in PISA. [b] Nation did not participate in TIMSS. [c] Data not available.

the mean of four items; students rated their agreement with statements such as "I am interested in the things I learn in Mathematics." Anxiety in Mathematics is assessed with the mean of five items; students rated their agreement with statements such as "I get very nervous doing Mathematics problems." Self-Concept in Mathematics is based on the mean of five items; students rated their agreement with statements such as "I learn Mathematics quickly." Self-Efficacy in Mathematics is composed of the mean of eight items; students rated how confident they felt about doing mathematics tasks by agreeing with statements such as "Solving an equation like $3x + 5 = 17$." Effect sizes for gender differences in achievement, attitudes, and affect on these scales are shown in Table 2.

Gender Equity Indicators

Composite indicators. Key to testing the gender stratification hypothesis is the valid assessment of gender equity on the national level. Given the theoretical emphasis on economic, educational, and political opportunities being greater for males, our measurement of gender equity should weight these dimensions more heavily than the dimension of health. Four of the most commonly used and well-regarded composite indicators of gender equity were chosen for the current study. When possible, data were obtained from *HDR 2003* (UNDP, 2003). Although the *HDR* is published annually, data from the 2003 report were chosen because they were collected during the same time period as the TIMSS and PISA data used in the meta-analysis. The UNDP sample, which includes up to 180 nations, includes 41 of the 46 TIMSS and 38 of the 41 PISA nations used in the meta-analysis; however, complete data were not available for all nations. Values for all composite indices, by nation, appear in Table 3. Intercorrelations among all composite indices of gender equity are shown in Table 4.

GEM. The GEM was developed by the UNDP (1995) for the purpose of assessing women's empowerment in political and economic spheres. It is composed of three factors: the extent of women's political participation and decision making (as measured by women's and men's percentage shares of parliamentary seats), economic participation and decision-making power (as measured by women's and men's percentage shares of positions as legislators, senior officials, and managers and women's and men's percentage shares of professional and technical positions), and the power exerted by women over economic resources (as measured by women's and men's estimated earned income). Higher values on the GEM indicate greater female empowerment. For further details on calculating the GEM, see UNDP (1995, 2003). It is notable that, although the GEM emphasizes the economic and political dimensions of gender equity, it omits the education dimension.

GEQ. The GEQ (White, 1997) was designed to tap into the implicit or underlying index of gender equality in the GDI (UNDP, 1995). Not a measure of gender equity as such, the GDI is a variation of the Human Development Index (HDI), which is a composite indicator of the areas of education (enrollment and literacy rates), health (life expectancy), and earned income. The GDI assesses the same components as the HDI but discounts those components for their gender disparities. Because the GDI is often mistaken and misused as an indicator of gender equality, several alternative indicators composed of the same elements (education, health, and earned income) have been proposed (Dijkstra, 2006; Schüler, 2006). The GEQ is one such alternative; it is simply the ratio of the GDI to the HDI. The GEQ's equal weighting of the education, economic, and health dimensions is a limitation in the context of a focus on math achievement. It was therefore predicted to correlate less strongly with gender differences in math performance than would other gender equity measures.

SIGE. The SIGE (Dijkstra, 2002) was developed in response to a workshop held at The Hague, in which social scientists identified eight dimensions of gender equality that could be observed across many different cultures. The SIGE attempts to build upon the GDI and GEM by using some of the components of those composite indices. Five variables make up the SIGE: (a) relative female-to-male access to education (1/3 combined primary, secondary, and tertiary enrollment; 2/3 adult literacy); (b) relative female-to-male life expectancy; (c) relative female-to-male labor market participation or economic activity rate; (d) women's share in higher labor market occupations (e.g., technical, professional, administrative, and management positions); and (e) women's share in parliamentary seats. Each of the five variables is standardized and is computed as the difference of the arithmetic mean of scores on all nations on indicator *j* from the score of nation *i* on indicator

PISA											
N_F	N_M	Math	Quantity	Space	Change	Uncertainty	EM	IM	Anx.	MSC	MSE
2,876	2,360	−0.05	−0.05	0.05	−0.10	−0.07	−0.20	0.06	−0.11	0.28	0.13
2,395	2,326	0.15	0.18	0.18	0.11	0.09	0.17	0.27	−0.35	0.34	0.27
2,184	2,671	0.14	0.16	0.12	0.05	0.19	−0.06	0.10	−0.20	0.19	0.25
5,073	4,462	c	c	c	c	c	0.27	0.15	−0.39	0.41	0.39
2,705	2,750	0.07	0.04	0.16	0.06	0.03	0.10	0.16	−0.23	0.27	0.19
2,989	2,846	0.12	0.11	0.21	0.05	0.08	0.16	0.11	−0.21	0.24	0.31

j, divided by the standard deviation of scores of all nations on indicator *j*. Depending on skew and kurtosis, variables may undergo transformation to normalize their distributions (as was done for the education variable). The SIGE places relatively more emphasis on the economic dimension than on the educational, health, and political dimensions.

GGI. The GGI (Hausmann et al., 2007), published by the World Economic Forum, was designed to assess gender inequities in access to resources and opportunities. As with the other GDI and GEM alternatives, it does not take into account overall human development of a nation and is not framed as a measure of women's empowerment. It is composed of four subindices, including economic participation and opportunity, educational attainment, political empowerment, and health and survival. The economic subindex is based on female-to-male ratios in five areas: labor force participation; wage equality; earned income; share of legislators, senior officials, and managers; and share of professional and technical positions. The education subindex is based on female-to-male ratios in four areas: literacy, primary enrollment, secondary enrollment, and tertiary enrollment. The political subindex is based on the female-to-male ratios in parliamentary seats and ministerial positions and the number of years during the past 50 with a female head of state. The health subindex includes the female-to-male life expectancy ratio and the sex ratio at birth. All but the health ratios are truncated at the "equality benchmark," which is set at 1. Next, the subindices are computed using the weighted mean of the ratios. The GGI is computed using the unweighted mean of the four subindices. For the current study, we used the 2007 GGI values, which appear in the Global Gender Gap Report 2007 and are based on 2005 data. For further details on computation of the GGI, see Hausmann et al. (2007). Because the GGI is the unweighted average of the four subindices, the educational, economic, political, and health dimensions contribute equally to a nation's gender equity assessment.

Of the four composite indices of gender equity used in the current study, the GEM is perhaps the most theoretically relevant, as it is composed only of economic and political dimensions. However, its exclusion of the educational dimension is a notable shortcoming. Although the GEQ, SIGE, and GGI all include the education dimension, which is fundamental to the gender stratification hypothesis, these composites include the health dimension, which is theoretically less relevant. Because of these limitations in the composite indices, we chose seven domain-specific indicators that were used in the computation of the above composite indices of gender equity and were theoretically relevant to the gender stratification hypothesis to assess the relationship between the gender gap in math and specific domains of gender equity.

Domain-specific indicators. The seven domain-specific indicators encompass education, political, and economic domains of gender equity. The education domain includes (a) net primary school enrollment, (b) net secondary school enrollment, and (c) gross tertiary school enrollment, which are calculated as ratio of the number of female students enrolled in that level of education as a percentage of the female population of official school age for that level to the number of male students enrolled in that level of education as a percentage of the male population of official school age for that level. The political domain is assessed by women's share (as percentage of total) of parliamentary seats. The economic domain includes (a) the female-to-male ratio in economic activity rate (the ratio of the proportion of females age 15 or older who contribute or are available to contribute to the production of goods and services to the proportion of males age 15 or older who contribute or are available to contribute to the production of goods and services); (b) women's share (as percentage of total) of higher labor market positions (technical and professional, as well as administrative and management positions); and (c) women's share (as percentage of total) of research positions (according to the International Standard Classification of Occupations [International Labour Organization, 1990]). When possible, data for domain-specific indicators (except women's share of research positions) were obtained from *HDR 2003* (UNDP, 2003); due to improvements in statistical reporting among some nations, some data were available only in later editions of the *HDR*. Data for the indicator of women's share of research positions refer to the year 2003 and were obtained from UNESCO Institute for Statistics (2006) and OECD.Stat (http://stats.oecd.org/). Values for each of the domain-specific gender equity indicators, by nation, appear in Table 3. Intercorrelations among domain-specific and composite indicators of gender equity appear in Table 4.

Table 3
Composite and Domain-Specific Indicators of Gender Equity for Each Nation

Nation	GEM	GEQ	SIGE	GGI	Prim.	Sec.	Tert.	EACT	HLMP	WR	Parl.
Armenia	[a]	0.997	0.149	0.665	1.02	1.06	1.25	0.88	[a]	0.46	0.03
Australia	0.754	0.999	0.296	0.720	1.01	1.03	1.24	0.77	0.35	[a]	0.27
Austria	0.782	0.995	0.312	0.706	1.01	0.99	1.14	0.65	0.39	0.21	0.31
Bahrain	[a]	0.988	−1.008	0.593	1.01	1.07	1.59	0.39	[a]	[a]	0.06
Belgium	0.695	0.994	0.202	0.720	1.00	0.97	1.13	0.66	0.35	0.28	0.25
Botswana	0.564	0.995	−0.039	0.680	1.04	1.14	0.89	0.77	0.44	[a]	0.17
Brazil	0.490	0.991	0.452	0.664	0.93	1.08	1.29	0.52	0.62	0.46	0.09
Bulgaria	0.606	0.999	0.733	0.709	0.98	0.98	1.35	0.86	0.47	0.47	0.26
Canada	0.771	0.997	0.451	0.720	1.00	1.01	1.33	0.82	0.44	[a]	0.24
Chile	0.467	0.988	−0.505	0.648	0.99	0.76	0.92	0.49	0.37	0.33	0.10
Chinese Taipei	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	0.19	[a]
Cyprus	0.542	0.994	−0.473	0.652	1.01	1.02	1.29	0.62	0.31	0.31	0.11
Czech Republic	0.579	0.995	0.288	0.672	1.00	1.02	1.05	0.83	0.40	0.28	0.16
Denmark	0.825	0.998	0.609	0.752	1.00	1.03	1.35	0.84	0.36	0.28	0.38
Egypt	0.253	0.978	−1.633	0.581	0.95	0.96	[a]	0.45	0.20	[a]	0.02
England	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
Estonia	0.560	0.998	1.043	0.701	0.98	1.03	1.55	0.82	0.53	0.43	0.18
Finland	0.801	0.998	0.937	0.804	1.00	1.02	1.21	0.87	0.43	0.30	0.37
France	[a]	0.998	0.283	0.682	1.00	1.02	1.23	0.77	0.42	0.30	0.12
Germany	0.776	1.003	0.325	0.762	1.02	1.01	0.96	0.70	0.39	0.12	0.31
Ghana	[a]	0.995	−0.680	0.673	0.95	0.86	0.40	0.98	[a]	[a]	0.09
Greece	0.519	0.993	−0.323	0.665	1.00	1.03	1.14	0.59	0.36	0.37	0.09
Hong Kong	[a]	0.997	−0.326	[a]	0.94	0.96	0.95	0.65	0.32	[a]	[a]
Hungary	0.518	0.996	0.381	0.673	0.99	1.01	1.27	0.71	0.48	0.35	0.10
Iceland	0.847	0.998	0.661	0.784	1.00	1.05	1.74	0.83	0.43	0.39	0.35
Indonesia	[a]	0.993	−0.709	0.655	0.99	0.96	0.77	0.68	[a]	[a]	0.08
Iran, Islamic Republic of	[a]	0.976	−1.464	0.590	0.98	0.94	0.93	0.38	0.25	[a]	0.04
Ireland	0.683	0.992	−0.187	0.746	1.00	1.04	1.27	0.53	0.39	0.31	0.14
Israel	0.612	0.994	−0.169	0.697	1.00	1.01	1.39	0.68	0.41	[a]	0.15
Italy	0.561	0.993	−0.252	0.650	1.00	1.01	1.32	0.59	0.32	0.29	0.10
Japan	0.515	0.994	−0.284	0.646	1.00	1.01	0.85	0.67	0.27	0.12	0.10
Jordan	[a]	0.981	−1.352	0.620	1.01	1.07	1.14	0.35	[a]	0.18	0.03
Korea, Republic of	0.363	0.993	−0.531	0.641	1.01	1.00	0.59	0.70	0.20	0.11	0.06
Latvia	0.576	0.999	1.151	0.733	1.00	1.08	1.65	0.80	0.53	0.53	0.21
Lebanon	[a]	0.980	−1.020	[a]	1.00	1.09	1.09	0.39	[a]	[a]	0.02
Liechtenstein	[a]	[a]	0.012	[a]	1.01	0.87	0.37	[a]	[a]	[a]	0.24
Lithuania	0.499	0.999	0.883	0.723	0.99	1.01	1.51	0.80	0.58	0.48	0.11
Luxembourg	[a]	0.989	0.000	0.679	1.01	1.08	1.24	0.58		0.17	0.17
Macao	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	0.11	[a]
Macedonia	[a]	1.014	−0.089	0.697	1.00	0.98	1.32	0.72	0.41	0.48	0.18
Malaysia	0.503	0.992	−0.439	0.644	1.00	1.11	1.08	0.61	0.33	0.34	0.15
Mexico	0.516	0.988	−0.436	0.644	1.01	1.08	0.96	0.48	0.33	0.32	0.16
Moldova, Republic of	0.468	0.996	0.546	0.717	1.00	1.03	1.29	0.84	0.52	0.30	0.13
Morocco	[a]	0.974	−1.511	0.568	0.91	0.83	0.80	0.52	0.24	0.27	0.06
Netherlands	0.794	0.996	0.260	0.738	0.99	1.00	1.07	0.67	0.37	0.17	0.33
New Zealand	0.750	0.997	0.619	0.765	1.00	1.02	1.52	0.80	0.46	0.39	0.29
Norway	0.837	0.997	0.699	0.806	1.00	1.01	1.52	0.85	0.37	0.29	0.36
Palestinian National Authority	[a]	[a]	−1.685	0.763	1.02	1.08	0.96	0.13	0.22	[a]	[a]
Philippines	0.539	0.996	0.210	[a]	1.01	1.18	1.10	0.61	0.60	0.55	0.17
Poland	0.594	0.998	0.683	0.676	1.00	1.03	1.44	0.80	0.46	0.39	0.21
Portugal	0.647	0.996	0.374	0.696	0.96	1.08	1.37	0.72	0.41	0.44	0.19
Romania	0.460	0.997	0.215	0.686	0.99	1.02	1.20	0.76	0.43	0.43	0.10
Russian Federation	0.440	0.994	1.035	0.687	1.00	0.99	1.36	0.82	0.51	0.43	0.06
Saudi Arabia	[a]	0.966	−1.830	0.565	0.92	0.95	1.29	0.28	0.19	0.17	0.00
Scotland	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
Serbia	[a]	[a]	0.598	[a]	1.00	1.01	1.20	0.76	0.71	0.43	0.20
Singapore	0.594	0.995	−0.536	0.661	[a]	[a]	[a]	0.64	0.34	0.26	0.12
Slovak Republic	0.598	0.998	0.622	0.680	1.01	1.01	1.09	0.84	0.46	0.41	0.19
Slovenia	0.582	0.998	0.368	0.684	0.99	[a]	1.35	0.81	0.43	0.34	0.12
South Africa	[a]	0.991	0.443	0.719	0.98	1.12	1.23	0.59	[a]	0.35	0.30
Spain	0.709	0.993	0.245	0.744	1.01	1.03	1.15	0.57	0.39	0.36	0.27
Sweden	0.831	0.999	1.071	0.815	0.99	1.04	1.52	0.89	0.40	0.35	0.45
Switzerland	0.720	0.995	0.000	0.692	0.99	0.95	0.78	0.66	0.34	0.21	0.22

(*table continues*)

Table 3 (*continued*)

Nation	GEM	GEQ	SIGE	GGI	Prim.	Sec.	Tert.	EACT	HLMP	WR	Parl.
Thailand	0.457	0.997	0.206	0.682	0.97	1.05	0.82	0.85	0.41	0.46	0.10
Tunisia	[a]	0.982	−1.141	0.628	0.99	1.05	0.97	0.48	[a]	[a]	0.12
Turkey	0.290	0.989	−1.173	0.577	0.95	0.82	0.70	0.62	0.20	0.36	0.04
United Kingdom	0.675	0.998	0.186	0.744	1.00	1.02	1.27	0.74	0.37	[a]	0.17
United States	0.760	0.998	0.446	0.700	1.01	1.02	1.32	0.82	0.50	[a]	0.14
Uruguay	0.516	0.995	0.345	0.661	1.01	1.11	1.83	0.67	0.45	0.47	0.12

Note. On all indicators, higher values reflect higher status of women. GEM = Gender Empowerment Measure; GEQ = Gender Equality Index; SIGE = Standardized Index of Gender Equality; GGI = Gender Gap Index; Prim. = male/female primary enrollment ratio; Sec. = male/female secondary enrollment ratio; Tert. = male/female tertiary enrollment ratio; EACT = male/female economic activity rate ratio; HLMP = women's share of higher labor market positions; WR = women's share of research positions; Parl. = women's share of parliamentary seats.
[a] Data not available.

Data Analysis

To estimate the magnitude of gender differences in mathematics achievement across countries, we conducted meta-analysis on each composite and content domain of the TIMSS and PISA data sets, using the formulas provided by Hedges and Becker (1986) and Lipsey and Wilson (2001). We also analyzed cognitive domains from the TIMSS data set using these methods.

Effect size computation. We computed effect sizes of gender differences for each of the 11 measures of mathematics achievement and the seven measures of mathematics attitudes and affect. The effect size, *d,* is defined as the mean for males minus the mean for females, divided by the pooled within-gender standard deviation. Thus, positive values of *d* represent higher scores for males than females, whereas negative values represent higher scores for females. Cohen (1988) provided guidelines for the interpretation of effect sizes; effect sizes of $d = 0.20$, 0.50, and 0.80 are considered small, medium, and large, respectively. We characterize effect sizes of $d < 0.10$ as negligible or close to zero (Hyde, 2005). All effect sizes were corrected for bias, using the formula provided by Hedges and Becker (1986). Analyses were conducted using both weighted and unweighted effect sizes; although results using unweighted effect sizes were very similar, we report the results using weighted effect sizes here. Meta-analytic results based on unweighted effect sizes are available from Nicole M. Else-Quest. Computed, unweighted effect sizes, in addition to male and female sample sizes, for each country of the TIMSS and PISA samples appear in Table 2.

Mixed-effects models and moderator analyses. Opinion about the best statistical methods for estimation of the homogeneity (Q) of effect sizes and testing moderators is in a state of flux. Meta-analyses have traditionally been based on fixed-effects models, which assume that variability among effect sizes is completely systematic and accounted for by the moderators in the analysis. In recent years, some have chosen instead to compute homogeneity statistics using the random-effects model, which assumes that variability among effect sizes is random (Hedges & Vevea, 1998). Whereas the fixed-effects model requires untenable statistical assumptions about the homogeneity of the sample of effect sizes (thereby increasing the Type I error rate), the random-effects model enlarges the error term so greatly that it is difficult to find significant moderators (such as indicators of gender equity), even with large samples of studies (Lipsey & Wilson, 2001). In fact, the random-effects analysis often produces nonsignificant homogeneity, which implies that the fixed-effects model could have been appropriate. For these reasons, we conducted the current meta-analysis using the mixed-effects model (Lipsey & Wilson, 2001), which attributes effect-size variability to systematic between-study variations, subject-

Table 4
Intercorrelations Between Composite and Domain-Specific Indicators of Gender Equity

Variable	1	2	3	4	5	6	7	8	9	10	11
1. GEM	—										
2. GEQ	.59**	—									
3. SIGE	.54**	.77**	—								
4. GGI	.84**	.69**	.72**	—							
5. Prim.	.34*	.47**	.30*	.41**	—						
6. Sec.	.17	.22	.29*	.32*	.45**	—					
7. Tert.	.37*	.26*	.51**	.39**	.14	.47**	—				
8. EACT	.38**	.77**	.78**	.49**	.18	−.01	.18	—			
9. HLMP	.17	.55**	.78**	.46**	.22	.42**	.52**	.53**	—		
10. WR	−.26	.38**	.47**	.15	−.06	.27	.48**	.37**	.68**	—	
11. Parl.	.88**	.52**	.65**	.84**	.30*	.16	.26*	.48**	.23	−.04	—

Note. Pairwise $n = 40–64$. GEM = Gender Empowerment Measure; GEQ = Gender Equality Index; SIGE = Standardized Index of Gender Equality; GGI = Gender Gap Index; Prim. = male/female primary enrollment ratio; Sec. = male/female secondary enrollment ratio; Tert. = male/female tertiary enrollment ratio; EACT = male/female economic activity rate ratio; HLMP = women's share of higher labor market positions; WR = women's share of research positions; Parl. = women's share of parliamentary seats.
* $p < .05$. ** $p < .01$.

level sampling error, and random effects. In the mixed-effects model, total homogeneity (Q_T) is computed using random-effects variance components, which are based on the residual homogeneity (Q_W) computed after taking moderators into account.

Moderator analyses tested the gender stratification hypothesis. We ran hierarchical multiple regressions, using appropriate weighting and mixed-effects model estimates, between the four composite indices (i.e., GEM, GEQ, SIGE, GGI) and seven domain-specific indicators (i.e., primary enrollment ratio, secondary enrollment ratio, tertiary enrollment ratio, economic activity rate ratio, women's share of higher labor market positions, women's share of research positions, and women's share of parliamentary seats) of gender equity and variables of gender differences in mathematics achievement, attitudes, and affect. The gender stratification hypothesis predicted that higher levels of gender equity in all domains would predict smaller gender differences in mathematics achievement, attitudes, and affect.

Results

Estimating the Magnitude and Variability of the Gender Gap

TIMSS.

Achievement. Table 5 displays mean weighted effect sizes (d), 95% confidence intervals, sample sizes (k), and homogeneity statistics (Q) based on the TIMSS-Math and TIMSS content domains of Algebra, Data, Geometry, Measurement, and Number. The weighted mean effect size of the gender difference in performance on the TIMSS-Math across the $k = 46$ effect sizes was $d = -0.01$, indicating that boys and girls performed similarly overall. This effect size reflects a gender difference of less than 1 point on the TIMSS-Math, which had an international average score of 467. The range of effect sizes for the TIMSS-Math was from $d = -0.42$ (Bahrain) to $d = 0.40$ (Tunisia), with 63.0% of the effect sizes smaller than $d = 0.10$. Small effect sizes also were observed in Data, Geometry, and Number. Although the gender difference in Measurement was significant ($d = 0.07$, $p < .05$), it is so small that it can be considered negligible. Girls outperformed boys slightly but significantly ($p < .05$) in Algebra. According to content domain, 39.1% of the effect sizes in Algebra, 56.5% of those in Data, 60.9% of those in Geometry, 58.7% of those in Measurement, and 54.3% of those in Number can be classified as negligible ($d < 0.10$; see Table 2). These results indicate a range of effect sizes distributed around a very small mean gender difference in math achievement on the TIMSS.

In addition, gender differences in achievement on the three cognitive domains were meta-analyzed. The mean weighted effect sizes for Knowing ($d = -0.04$, $p > .05$), Applying ($d = 0.02$, $p > .05$), and Reasoning ($d = -0.05$, $p < .05$) were all negligible in magnitude but significantly nonhomogeneous. Thus, in contrast to predictions based on the greater male variability hypothesis, there is little evidence that gender differences vary according to the difficulty level of the TIMSS items.

Attitudes and affect. Table 5 displays mean weighted effect sizes (d), 95% confidence intervals, sample sizes (k), and homogeneity statistics (Q) based on the TIMSS scales of Self-Confidence in Math and Valuing Math. The weighted mean effect size of the gender difference in self-confidence was $d = 0.15$, indicating that, averaged across nations, males reported higher self-confidence in math than did

Table 5
Weighted Mean Effect Sizes (d), 95% Confidence Intervals (95% CI), Number of Effect Sizes (k), and Homogeneity Statistics (Q_T) for Gender Differences in Math Achievement, Attitudes, and Affect

Content domain	d	k	95% CI	Q_T
		TIMSS		
Math	−0.01	46	[−0.05, 0.03]	46.42
Algebra	−0.11**	46	[−0.15, −0.07]	45.61
Data	0.00	46	[−0.04, 0.04]	47.45
Geometry	0.01	46	[−0.03, 0.05]	46.90
Measurement	0.07**	46	[0.04, 0.11]	46.13
Number	0.01	46	[−0.03, 0.05]	45.70
Self-Confidence in Math	0.15**	46	[0.12, 0.19]	46.00
Valuing Math	0.10**	46	[0.07, 0.14]	44.22
		PISA		
Math	0.11**	40	[0.09, 0.13]	51.31
Quantity	0.06**	40	[0.04, 0.09]	53.19
Space/Shape	0.15**	40	[0.13, 0.18]	52.03
Change/Relationships	0.09**	40	[0.07, 0.11]	45.98
Uncertainty	0.12**	40	[0.09, 0.14]	40.49
Extrinsic Motivation	0.24**	41	[0.18, 0.29]	48.46
Intrinsic Motivation	0.20**	41	[0.17, 0.24]	47.99
Anxiety	−0.28**	41	[−0.31, −0.25]	47.66
Self-Concept	0.33**	41	[0.28, 0.36]	48.81
Self-Efficacy	0.33**	41	[0.30, 0.37]	43.81

Note. TIMSS = Trends in International Mathematics and Science Study; PISA = Programme for International Student Assessment.
* $p < .05$. ** $p < .01$.

girls by a small amount. Effect sizes for the self-confidence in math scale ranged from $d = -0.12$ (Bahrain) to $d = 0.43$ (Hong Kong), with 30.4% of the effect sizes smaller than $d = 0.10$. Similarly, the weighted mean effect size of the gender difference in valuing math was $d = 0.10$. The range of effect sizes for the valuing math scale was from $d = -0.16$ (Philippines) to $d = 0.46$ (Netherlands), with 34.8% of the effect sizes smaller than $d = 0.10$. The mean effect sizes in self-confidence and valuing math are small but statistically significant ($p < .05$). Gender differences in math achievement were significantly correlated with gender differences in self-confidence in math ($r = .54$, $p < .01$) and students' valuing mathematics ($r = .30$, $p < .05$).

PISA.

Achievement. See Table 5 for mean weighted effect sizes (d), 95% confidence intervals, sample sizes (k), and homogeneity statistics (Q) based on the PISA-Math and PISA content domains of Quantity, Space/Shape, Change/Relationships, and Uncertainty. The weighted mean effect size of the gender difference in performance on the PISA-Math across the $k = 40$ effect sizes was $d = 0.11$, indicating that boys performed slightly better than girls overall. This effect size indicates a gender difference of 11 points on the PISA-Math, which had an average OECD score of 500. A similar pattern of gender differences was observed for the four PISA content domains: Boys outperformed girls, but the gender differences are negligible to very small in magnitude. The effect sizes for each nation (shown in Table 2) demonstrate that the majority (50%) of the effect sizes favor boys, 2.5% favor girls, and 45% are negligible ($d < 0.10$). The range of effect sizes for the PISA-Math was from $d = -0.17$ (Iceland) to $d = 0.29$ (Liechtenstein). According to content domain, 60.0% of the effect sizes in Quantity, 17.5% of those in Space/Shape, 47.5% of those in Change/Relationships, and 42.5% of those in Uncertainty can be classified as negligible. These results indicate cross-national variation in the magnitude of the gender gap in math achievement, with males outperforming females by a small amount on average.

Attitudes and affect. Table 5 shows mean weighted effect sizes (d), 95% confidence intervals, sample sizes (k), and homogeneity statistics (Q) for the PISA scales of Extrinsic Motivation, Intrinsic Motivation, Anxiety, Self-Efficacy, and Self-Concept. Although mean effect sizes are small, boys consistently reported significantly more positive attitudes about mathematics ($d = 0.21$–0.33). Similarly, girls reported greater mathematics anxiety than did boys ($d = -0.27$). Ranges of gender differences in math attitudes and affect were larger than those in math achievement. In extrinsic motivation, gender differences ranged from $d = -0.20$ (Thailand) to $d = 0.89$ (Liechtenstein), with 76.9% of them favoring males (20.5% were negligible); for intrinsic motivation, gender differences ranged from $d = 0.01$ (Russian Federation) to $d = 0.60$ (Liechtenstein), with 82.1% of them favoring males (17.9% were negligible). Similarly, gender differences in self-concept ranged from $d = 0.07$ (Russian Federation) to $d = 0.77$ (Liechtenstein), with 97.4% of them favoring males (2.6% were negligible); for self-efficacy, gender differences ranged from $d = 0.08$ (Indonesia) to $d = 0.65$ (Liechtenstein), with 97.4% of them favoring males (2.6% were negligible). Gender differences in anxiety ranged from $d = -0.61$ (Liechtenstein) to $d = -0.03$ (Poland), with 94.9% of them showing higher anxiety among girls (5.1% were negligible). In contrast to the results in mathematics achievement, which showed gender similarities, mathematics attitudes and affect showed consistent gender differences, with boys reporting greater extrinsic and intrinsic motivation, self-concept, and self-efficacy and less math anxiety. Effect sizes in achievement, attitudes, and affect varied a great deal across nations. National gender differences in PISA-Math achievement were significantly correlated with national gender differences in extrinsic motivation ($r = .40$, $p < .05$), intrinsic motivation ($r = .35$, $p < .05$), math anxiety ($r = -.35$, $p < .05$), self-concept ($r = .33$, $p < .05$), and self-efficacy ($r = .33$, $p < .05$).

Comparing results from TIMSS and PISA. Eighteen nations participated in both TIMSS and PISA; we compared gender differences on the math composites from the two assessments. Gender differences on the two assessments were not significantly correlated in magnitude ($r = .27$, $p > .05$), though lack of statistical significance is probably due, in part, to the small sample of nations participating in both assessments. Paired-samples t tests demonstrated that the differences between the TIMSS-Math and PISA-Math assessments are not substantial (mean difference $= 0.05$, $SD = 0.11$, $t = 1.92$, $p > .05$). Figure 1 displays the effect sizes for the TIMSS and PISA math composites. As demonstrated, their distributions overlap significantly.

Items from the TIMSS scale of Self-Confidence in Math and the PISA scale of Self-Concept are very similar; both include items reflecting students' perceptions of their ability to do well in math and to learn math quickly. Thus, as with the math composites, we conducted analyses to compare the effect sizes in nations that participated in both assessments. Gender differences on the two assessments were highly correlated ($r = .77$, $p < .001$). Yet, paired-samples t tests demonstrated that gender differences on the two assessments differed significantly within nations (mean difference $= -0.08$, $SD = 0.09$, $t = -3.81$, $p < .001$). Figure 2 displays the distribution of effect sizes on the PISA scale of Self-Concept and the TIMSS scale of Self-Confidence in Math.

Moderator Analyses: The Gender Stratification Hypothesis

Composite indices of gender equity.

Achievement. What factors explain the cross-national variability in the magnitude and direction of gender differences in math? Multiple regression analyses tested the gender stratification hypothesis. Composite indices of gender equity (e.g., GEM, GEQ, SIGE, GGI) were tested as moderators of gender differences in math achievement, attitudes, and affect. Each of the four composite indices of gender equity was entered into regression analyses as an individual predictor of each of the two gender differences in math achievement (TIMSS and PISA math composites) and the seven gender differences in attitudes and affect. Cases were weighted with the inverse variance of the effect sizes, which includes the random-effects variance component, as recommended by Lipsey and Wilson (2001). Results for these models appear in Table 6. Examination of the regression coefficients indicates a pattern that provides mixed support for the gender stratification hypothesis. The composite indices of gender equity did not significantly predict gender differences in math achievement on the TIMSS. In contrast, each of the four gender equity composite indices significantly and negatively predicted gender differences in PISA achievement, as hypothesized. These ef-

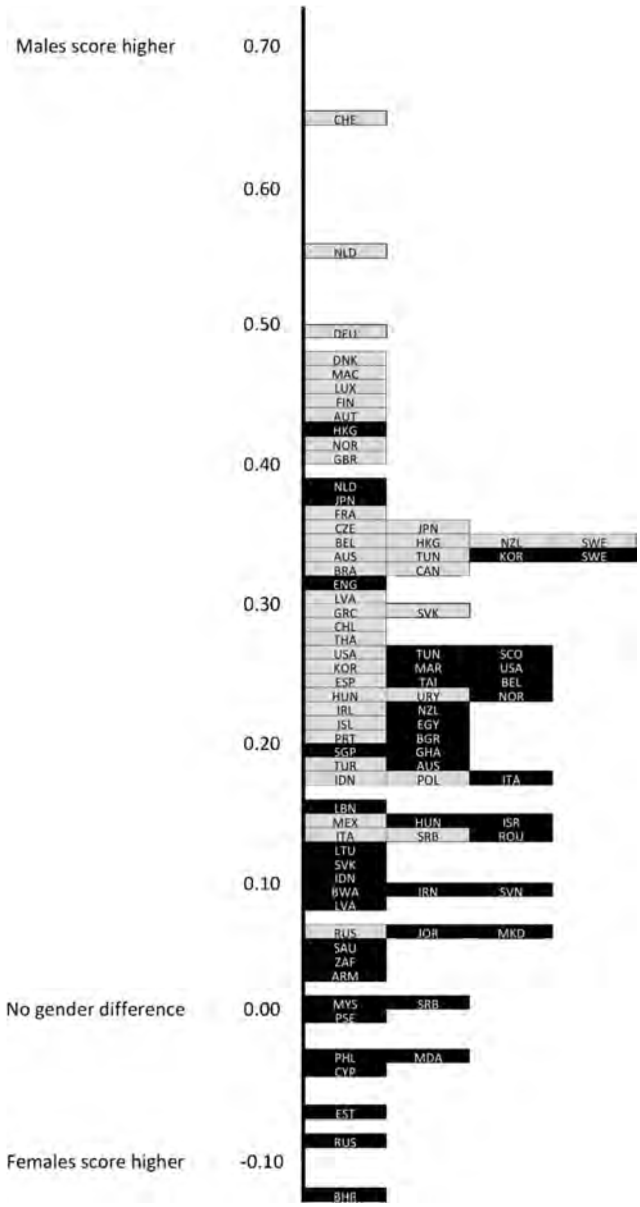

Figure 1. Distribution of effect sizes on TIMSS-Math (black) and PISA-Math (gray). Abbreviations for nations are as follows: ARM: Armenia; AUS: Australia; AUT: Austria; BEL: Belgium; BGR: Bulgaria; BHR: Bahrain; BRA: Brazil; BWA: Botswana; CAN: Canada; CHE: Switzerland; CHL: Chile; CYP: Cyprus; CZE: Czech Republic; DEU: Germany; DNK: Denmark; EGY: Egypt; ENG: England; ESP: Spain; EST: Estonia; FIN: Finland; FRA: France; GBR: United Kingdom; GHA: Ghana; GRC: Greece; HKG: Hong Kong; HUN: Hungary; IDN: Indonesia; IRL: Ireland; IRN: Islamic Republic of Iran; ISL: Iceland; ISR: Israel; ITA: Italy; JOR: Jordan; JPN: Japan; KOR: Republic of Korea; LBN: Lebanon; LIE: Liechtenstein; LTU: Lithuania; LUX: Luxembourg; LVA: Latvia; MAC: Macao; MDA: Republic of Moldova; MEX: Mexico; MKD: Macedonia, TFYR; MYS: Malaysia; NLD: Netherlands; NOR: Norway; NZL: New Zealand; PHL: Philippines; POL: Poland; PRT: Portugal; PSE: Palestinian National Authority; ROU: Romania; RUS: Russian Federation; SAU: Saudi Arabia; SCO: Scotland; SGP: Singapore; SRB: Serbia; SVK: Slovak Republic; SVN: Slovenia; SWE: Sweden; TAI: Chinese Taipei; THA: Thailand; TUN: Tunisia; TUR: Turkey; URY: Uruguay; USA: United States; ZAF: South Africa.

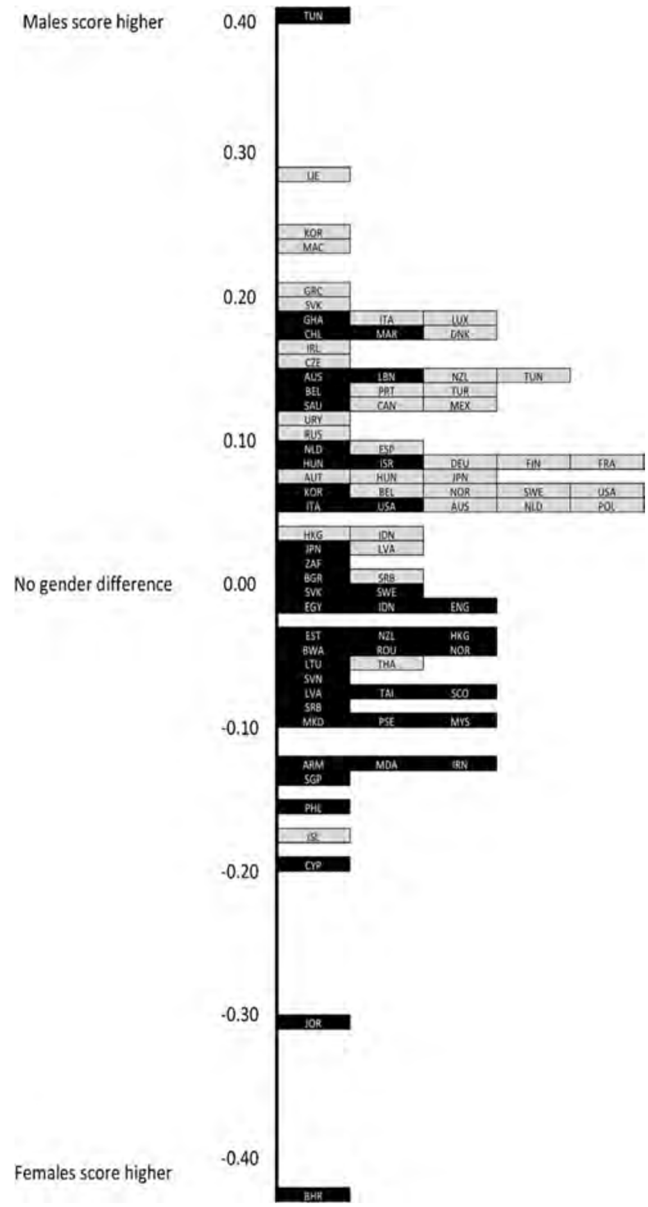

Figure 2. Distribution of effect sizes on TIMSS Self-Confidence in Mathematics (black) and PISA Self-Concept in Mathematics (gray). Abbreviations for nations are as follows: ARM: Armenia; AUS: Australia; AUT: Austria; BEL: Belgium; BGR: Bulgaria; BHR: Bahrain; BRA: Brazil; BWA: Botswana; CAN: Canada; CHE: Switzerland; CHL: Chile; CYP: Cyprus; CZE: Czech Republic; DEU: Germany; DNK: Denmark; EGY: Egypt; ENG: England; ESP: Spain; EST: Estonia; FIN: Finland; FRA: France; GBR: United Kingdom; GHA: Ghana; GRC: Greece; HKG: Hong Kong; HUN: Hungary; IDN: Indonesia; IRL: Ireland; IRN: Islamic Republic of Iran; ISL: Iceland; ISR: Israel; ITA: Italy; JOR: Jordan; JPN: Japan; KOR: Republic of Korea; LBN: Lebanon; LIE: Liechtenstein; LTU: Lithuania; LUX: Luxembourg; LVA: Latvia; MAC: Macao; MDA: Republic of Moldova; MEX: Mexico; MKD: Macedonia, TFYR; MYS: Malaysia; NLD: Netherlands; NOR: Norway; NZL: New Zealand; PHL: Philippines; POL: Poland; PRT: Portugal; PSE: Palestinian National Authority; ROU: Romania; RUS: Russian Federation; SAU: Saudi Arabia; SCO: Scotland; SGP: Singapore; SRB: Serbia; SVK: Slovak Republic; SVN: Slovenia; SWE: Sweden; TAI: Chinese Taipei; THA: Thailand; TUN: Tunisia; TUR: Turkey; URY: Uruguay; USA: United States; ZAF: South Africa.

Table 6
Variance in Gender Differences in Math Achievement, Attitudes, and Affect Explained by Composite Indices of Gender Equity (R^2), Weighted Multiple Regression Coefficients (β), Residual Homogeneity Statistics (Q_W), and Number of Effect Sizes (k)

Content domain	GEM	GEQ	SIGE	GGI
TIMSS				
Math Composite				
R^2	0.04	0.04	0.00	0.00
β	0.20	−0.19	−0.06	0.01
Q_W	27.80	41.02	43.04	39.04
k	28	41	43	39
Self-Confidence in Math				
R^2	0.11	0.00	0.00	0.02
β	0.33	0.01	−0.04	0.15
Q_W	27.94	41.03	43.03	39.06
k	28	41	43	39
Valuing Math				
R^2	0.16	0.00	0.00	0.02
β	0.40*	0.00	0.05	0.13
Q_W	27.91	41.00	43.00	39.03
k	28	41	43	39
PISA				
Math Composite				
R^2	0.11	0.14	0.09	0.14
β	−0.33*	−0.38*	−0.31*	−0.38*
Q_W	33.07	38.05	40.61	37.17
k	32	37	39	36
Extrinsic Motivation				
R^2	0.20	0.03	0.01	0.09
β	0.45**	0.16	0.10	0.30
Q_W	33.01	38.99	42.20	37.01
k	33	38	40	37
Intrinsic Motivation				
R^2	0.12	0.00	0.00	0.01
β	0.35*	0.07	−0.07	0.12
Q_W	32.99	37.97	42.42	36.98
k	33	38	40	37
Anxiety				
R^2	0.41	0.00	0.00	0.19
β	−0.64**	−0.04	0.01	−0.43**
Q_W	33.28	38.04	42.00	37.31
k	33	38	40	37
Self-Concept				
R^2	0.27	0.04	0.00	0.12
β	0.52**	0.19	0.02	0.35*
Q_W	33.00	38.02	42.84	37.06
k	33	38	40	37
Self-Efficacy				
R^2	0.16	0.04	0.03	0.11
β	0.40**	0.19	0.18	0.33*
Q_W	33.19	38.09	41.83	37.10
k	33	38	40	37

Note. TIMSS = Trends in International Mathematics and Science Study; PISA = Programme for International Student Assessment; GEM = Gender Empowerment Measure; GEQ = Gender Equality Index; SIGE = Standardized Index of Gender Equality; GGI = Gender Gap Index.
* $p < .05$. ** $p < .01$.

fects were medium in magnitude, controlling for human development, partial correlations between the gender equity composite indices and mean female scores on PISA, as well as between the gender equity composite indices and mean male scores on the PISA. The partial correlations with female scores ($r = -.05$ to .36) tended to be larger but did not differ significantly from correlations with mean male scores ($r = -.17$ to .27). This result suggested that the cross-national variation in the gender gap on the PISA that is explained by gender equity is probably not due solely to relative improvements in female scores across nations.

In sum, there is some support for the gender stratification hypothesis when composite indices of gender equity are used, but there is variability among the indices in their power to predict cross-national gender differences in math achievement on the TIMSS and PISA. These issues are addressed further in the Discussion.

Attitudes and affect. In terms of attitudes and affect across the TIMSS and PISA data, the GEM was a significant moderator in six out of seven models. However, the direction of these effects was contrary to predictions: In nations with greater gender equity, gender differences in valuing math, extrinsic motivation, intrinsic motivation, self-concept, and self-efficacy were larger (favoring males) than in nations with less gender equity. Similarly, the gender difference in math anxiety—which was higher in girls than in boys—was largest in nations with more gender equity. These effects were medium to large in magnitude; notably, the GEM (which assesses women's political and economic empowerment) explained 41% of the variance in gender differences in math anxiety. Although the GEQ and SIGE were not significant moderators of gender differences in attitudes and affect, we see the same unpredicted direction of significant effects in three of the seven models predicted by the GGI. These effects were medium in magnitude.

Domain-specific indicators of gender equity.

Achievement. We used multiple regression analysis in order to assess the specific domains in which gender equity was linked to the gender gaps in math. Each of the seven domain-specific indicators of gender equity was entered as a predictor of the two composite indices of gender differences in math achievement and the seven gender differences in math attitudes and affect. The results (including R^2, β, Q_W, and k) of these models appear in Table 7. Inspection of the results reveals two distinct patterns. First, gender differences in math achievement on the TIMSS were negatively associated with gender ratios in primary, secondary, and tertiary school enrollment, as predicted by the gender stratification hypothesis. This pattern suggests that, in nations where girls have more equal access to education, girls and boys tend to perform similarly on the TIMSS. Note that all children tested on the TIMSS were enrolled in formal schooling at the eighth grade. Yet, when girls are not educated at the same rate as the boys in their communities, the girls who are enrolled in school tend not to perform at the same level as their male classmates. The second pattern that is apparent is that gender differences in math achievement on the PISA were significantly and negatively predicted by women's shares of research positions and parliamentary seats, as well as by the male-to-female ratio in economic activity rates, as predicted by the gender stratification hypothesis. These findings suggest that, when women participate at the rate of men in the labor market (particularly in science jobs) and in national government, the gender gap in math achievement on the PISA is smaller.

Attitudes and affect. In terms of attitudes and affect, results followed a pattern similar to that for achievement. As predicted, gender differences on the TIMSS attitudes scale of Self-Confidence in Math were negatively predicted by secondary and tertiary enrollment ratios, women's share of research positions, and women's share of higher labor market positions; this indicated that gender equity in education and upper level employment (particularly science jobs) is reflected in smaller gender gaps in students' math self-confidence. The gender difference on the TIMSS scale of Valuing Math was significantly predicted by the secondary enrollment ratio and women's shares of higher labor market positions and research jobs, also as predicted.

Similarly, gender differences in the PISA scale of Extrinsic Motivation were significantly (and positively) predicted by male-to-female ratios in primary enrollment ratios. In contrast, gender differences in each of the PISA math attitudes and affect scales were significantly linked to women's share of parliamentary seats. However, as with the models testing composite gender equity indices, each of these significant links was in the opposite of the predicted direction. When women enjoy greater political representation, gender differences in motivation, interest, self-concept, and self-efficacy (which were higher in boys) are larger and gender differences in anxiety (which was higher in girls) are smaller. These effects were medium in magnitude.

In order to determine the combined contributions of domain-specific gender equity indicators on gender differences in math, we entered significant moderators simultaneously into regression models. Primary, secondary, and tertiary enrollment ratios explained 25.5% of the variance in the gender gap in math achievement on TIMSS. When entered into a regression model simultaneously, secondary and tertiary enrollment ratios, women's share of higher labor market positions, and women's share of research jobs explained 36.4% of the variance in gender differences in math self-confidence. The combined effect of secondary enrollment ratio, women's share of higher labor market positions, and women's share of research jobs on students' valuing of mathematics was also substantial ($R^2 = .35$). On the PISA-Math, economic activity rate ratios, women's share of research jobs, and women's parliamentary representation explained 26.1% of the variance in the gender gap for achievement. When entered into a regression model simultaneously, primary enrollment ratios, women's share of research jobs, and women's parliamentary representation explained 51.7% of the variance in gender differences in extrinsic motivation. The combined effect of tertiary enrollment ratios, women's share of research jobs, and women's parliamentary representation on intrinsic motivation was $R^2 = .41$. Women's share of research jobs and parliamentary representation explained 34.4% of the variance in gender differences in math anxiety, 49% of the variance in gender differences in self-concept, and 28% of the variance in self-efficacy.

In sum, gender equity in various domains appears to explain a substantial amount of cross-national variation in gender differences in math achievement, attitudes, and affect. Across both the TIMSS and PISA attitudes and affect scales, gender differences were significantly predicted by women's share of research positions, such that nations with greater participation of women in scientific research show smaller gender differences in self-confidence, valuing math, intrinsic and extrinsic motivation, math anxiety, self-efficacy, and self-concept.

Discussion

Does the magnitude of national gender differences on TIMSS and PISA warrant further publicity of a gender gap in math performance? We meta-analyzed these data sets and found evidence of gender similarities in mathematics achievement, despite considerable cross-national variability in the direction and magnitude of effects. Our findings are consistent with previous reports on gender similarities in math achievement in the United States (Hedges & Nowell, 1995; Hyde, Fennema, & Lamon, 1990; Hyde et al., 2008). Previous reports on the 2003 TIMSS and PISA by the

Table 7
Variance in Gender Differences in Mathematics Achievement and Attitudes Explained by Domain-Specific Indicators of Gender Equity (R^2), Weighted Multiple Regression Coefficients (β), Residual Homogeneity Statistics (Q_W), and Number of Effect Sizes (k)

Content domain	Prim.	Sec.	Tert.	EACT	HLM	WR	Parl.
			TIMSS				
Math Composite							
R^2	0.13	0.17	0.14	0.00	0.08	0.07	0.01
β	−0.36*	−0.41**	−0.37*	0.07	−0.29	−0.26	0.08
Q_W	41.98	41.06	40.98	43.05	34.93	31.01	41.04
k	42	41	41	43	35	31	41
Self-Confidence							
R^2	0.08	0.16	0.12	0.02	0.15	0.29	0.08
β	−0.28	−0.41**	−0.35*	0.12	−0.38*	−0.54**	0.29
Q_W	42.02	41.11	40.98	43.04	34.99	31.07	40.95
k	42	41	41	43	35	31	41
Valuing Math							
R^2	0.01	0.11	0.01	0.00	0.11	0.23	0.07
β	−0.08	−0.33*	0.12	0.02	−0.33*	−0.48**	0.26
Q_W	42.01	41.15	40.99	43.01	35.07	30.98	40.88
k	42	41	41	43	35	31	41
			PISA				
Math Composite							
R^2	0.01	0.01	0.08	0.13	0.08	0.10	0.09
β	0.08	−0.10	−0.29	−0.37*	−0.29	−0.32*	−0.30*
Q_W	40.74	40.45	39.79	38.92	36.15	33.85	39.90
k	39	39	39	38	35	33	38
Extrinsic Motivation							
R^2	0.11	0.05	0.04	0.01	0.02	0.37	0.15
β	0.33*	−0.22	−0.20	−0.07	−0.15	−0.61**	0.39**
Q_W	42.06	41.38	41.44	38.98	36.01	33.05	41.40
k	40	40	40	39	36	33	39
Intrinsic Motivation							
R^2	0.02	0.07	0.12	0.01	0.04	0.33	0.11
β	0.13	−0.27	−0.35*	−0.07	−0.19	−0.57**	0.34*
Q_W	42.29	41.31	41.07	38.93	35.94	32.56	
k	40	40	40	39	36	33	
Anxiety							
R^2	0.00	0.01	0.01	0.02	0.08	0.23	0.17
β	−0.05	0.11	0.10	0.14	0.28	0.48**	−0.42**
Q_W	41.95	41.54	41.59	39.07	36.04	32.88	41.31
k	40	40	40	39	36	33	39
Self-Concept							
R^2	0.01	0.03	0.08	0.00	0.07	0.36	0.21
β	0.09	−0.18	−0.27	0.00	−0.26	−0.60**	0.46**
Q_W	42.68	41.99	41.63	38.99	35.94	32.54	42.35
k	40	40	40	39	36	33	39
Self-Efficacy							
R^2	0.03	0.04	0.00	0.00	0.01	0.15	0.17
β	0.18	−0.21	−0.06	0.00	−0.10	−0.39*	0.41**
Q_W	41.40	40.83	41.30	39.01	36.12	33.03	40.67
k	40	40	40	39	36	33	39

Note. Prim. = male/female primary enrollment ratio; Sec. = male/female secondary enrollment ratio; Tert. = male/female tertiary enrollment ratio; EACT = male/female economic activity rate ratio; HLMP = women's share of higher labor market positions; WR = women's share of research positions; Parl. = women's share of parliamentary seats; TIMSS = Trends in International Mathematics and Science Study; PISA = Programme for International Student Assessment.
* $p < .05$. ** $p < .01$.

IEA and OECD have supported our conclusions with these data sets (Mullis et al., 2004; OECD, 2004). Compared to those for previous waves of data collection for these data sets, the average effect size has changed very little. For the 1982 SIMS, the gender difference in math achievement ($d = 0.03$) was small and negligible (Baker & Jones, 1993); on the 1995 TIMSS, the gender difference was $d = 0.08$ (Mullis, Martin, Fierros, Goldberg, & Stemler, 2000). These data indicate a pattern of gender similarities in mean math achievement over 2 decades. Comparison of d for each nation across the three IEA studies indicates that, for over half of all nations, the gender gap in math has remained near zero or has even decreased in magnitude to become negligible. For example, in the United States, $d = -0.01$ in 1982, $d = 0.05$ in 1995, and $d = 0.06$ in 2003.

The largest mean effect size in math achievement was in the content domain of Space/Shape on PISA ($d = 0.15$); this content domain assesses understanding of spatial relationships. The finding is consistent with historical evidence of gender differences in the spatial skill of mental rotation (Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995) and with the neglect of spatial skill instruction in schools (National Research Council, 2006). Recent research shows that small amounts of instruction can produce large increases in spatial skills (Sorby, 2001; Sorby & Baartmans, 2000; Uttal, Hand, & Newcombe, 2009). Today, children have increasingly more computer experience, particularly in the form of playing videogames, which can improve mental rotation skills (Okagaki & Frensch, 1994) and in some cases eliminate gender differences in mental rotation (Saccuzzo, Craig, Johnson, & Larson, 1996). Furthermore, some research indicates that gender differences in videogame experiences mediate the gender difference in mental rotation (Terlecki & Newcombe, 2005).

Seven of the 11 mean effect sizes in math achievement were significant, though each was very small in magnitude; this provides further support for the gender similarities hypothesis (Hyde, 2005). Considerable cross-national variability in effect size magnitude and direction qualifies this finding. Although most national effect sizes were very small or negligible, some were small or medium, with males outperforming females in some cases and females outperforming males in others. These findings indicate that the gender gap in math persists in some nations but not in others.

Despite overall similarities in achievement, boys felt more confident and less anxious in their math abilities and were more extrinsically and intrinsically motivated to do well in math than were girls, as consistent with previous research findings (Casey et al., 1997; Fredricks & Eccles, 2002; Hyde, Fennema, Ryan, et al., 1990; McGraw et al., 2006; Meece et al., 1990; Pajares & Miller, 1994; Pomerantz, Altermatt, & Saxon, 2002). Boys also scored one third of a standard deviation higher than girls on math self-concept and self-efficacy; these were the largest mean effect sizes in the meta-analysis. The intuitively contradictory finding of gender similarities in achievement and gender differences in attitudes and affect is consistent with previous research (Hyde, Fennema, & Lamon, 1990; Hyde, Fennema, Ryan, et al., 1990; Pomerantz et al., 2002) and may be influenced by a variety of factors. Pomerantz et al. (2002) explained this pattern of results with U.S. students as being rooted in girls' greater concern for adult approval. That is, girls work hard to achieve in academics in order to please adults, but they also feel internal distress about disappointing those adults when they experience or anticipate academic failure. Because boys are less concerned with garnering adult approval, they are less motivated to achieve and are not deeply distressed by failing to do so. We must be cautious in generalizing such a pattern to other cultures. Other theoretical approaches are also informative. Consistent with Steele's (1997) theory of stereotype threat, gender stereotypes about gender roles and math may encourage girls to feel anxious and less confident (Steele, 1997). Alternatively, Roberts (1991) argued that girls and boys may approach such achievement situations differently. Whereas boys, feeling self-confident and capable, may respond positively to the competitive nature of achievement testing, girls, feeling anxious, may respond negatively to the personally evaluative nature of the situation (Roberts, 1991).

The link between math achievement and attitudes is complex. Although math achievement and attitudes are positively correlated on the student level, they are strongly and negatively correlated on the nation level (Shen & Tam, 2008). That is, nations with high math achievement tend to have students with less positive math attitudes. These nation-level data may reflect cultural standards for math achievement; in nations where math achievement is highly valued and where there is access to high-level math classes, students perceive math as being difficult and less enjoyable. For example, Japanese students outperform German students in math achievement but hold higher standards for themselves and feel more negatively about math (Randel, Stevenson, & Witruk, 2000). This pattern likely feeds back to the effort expended, such that students work hard to meet the high cultural standards set for them and feel anxious about their achievement. It is unclear whether a similar pattern explains gender differences in achievement and attitudes.

The Gender Stratification Hypothesis

The current study supported the gender stratification hypothesis in many analyses. In contrast to those of previously published reports (Baker & Jones, 1993; Guiso et al., 2008; Riegle-Crumb, 2005), our results point to specific domains of gender equity that may be directly or indirectly responsible for gender gaps in math. Results from models using composite indicators of gender equity were mixed. The gender differences on the PISA math composite showed some association with composite indices of gender equity; the GEM and GGI were the best composite predictors of the gender gap in math achievement, attitudes, and affect on PISA. In contrast, the TIMSS math composite showed no association with the composite indices of gender equity. Analyses of specific domains of gender equity are more revealing and indicate that gender equity in education and research jobs is most relevant.

The gender differences in math achievement and attitudes on TIMSS were consistently associated with gender ratios in school enrollment. This is perhaps not surprising, in light of the IEA's focus on classroom teaching and student learning. That is, because of its emphasis on assessing intended and attained curricula, TIMSS should be more sensitive to variations in the characteristics of curricula or educational institutions. In contrast, with PISA's emphasis on practical application and literacy, variations in social, economic, or cultural forces should be more relevant to test performance. Overall, these findings suggest that, consistent with the gender stratification hypothesis, gender equity in education is important not only for girls' math achievement but also for girls' self-confidence and valuing of mathematics. If girls recognize that they have the same rights to formal education that their male peers do, they may feel it is appropriate to work hard and invest in their education. Alternatively, if it is apparent that the education of girls is not highly valued in their community, it is likely that girls will not value their own educational achievements and will withdraw from achievement opportunities. If girls are outnumbered by boys in the classroom, their identity as girls becomes more salient and the risk for deleterious stereotype-threat effects is increased. Such arguments may be used to support gender-segregated education, insofar as a single-sex classroom removes gender from the learning process. Yet, evidence for such claims is mixed. Using TIMSS 2003 data, Wiseman (2008) found that nations with high percentages of gender-segregated math classrooms displayed considerable variability in gender differences in math achievement, whereas nations with few or no gender-segregated math classrooms displayed a pattern of gender similarities in achievement. Although causal inferences are unwarranted with the current data, when considered from the theoretical perspectives discussed herein, our results suggest that improving girls' access to formal education will not only increase the number of girls being educated but may also be a factor in higher achievement of the girls who are enrolled. Future research should focus on examining the psychological mechanisms responsible for the link between gender equity in enrollment and gender differences in math attitudes and achievement, as described by Eccles's expectancy-value theory, Bandura's social cognitive theory, and Eagly and Wood's social structural theory.

Gender equity in research jobs was also a consistent and strong predictor of the gender gap in math achievement, attitudes, and affect. This gender equity indicator comes closest to assessing the cultural mechanisms suggested by the gender stratification hypothesis and our theoretical foundation, which rests on psychological mechanisms proposed by social structural theory, expectancy-value theory, and cognitive social learning theory. That is, when girls develop in a societal context where women have careers in scientific research, they receive a clear message that STEM is within the realm of possibilities for them. Conversely, if girls' mothers, aunts, and sisters do not have STEM careers, they will perceive that STEM is a male domain and thus feel anxious about math, lack the confidence to take challenging math courses, and underachieve on math tests. Consistent with cognitive social learning theory and expectancy-value theory, girls need to observe women engaging in STEM in order to perceive STEM as a viable option for themselves. These findings provide substantive support for the gender stratification hypothesis and direct our attention to specific mechanisms responsible for gender gaps in math achievement, attitudes, and affect.

The gender stratification hypothesis maintains that indicators of societal gender equity predict gender differences in achievement, attitudes, and affect because the observed lower social status of women leads girls to achieve less and think and feel negatively about math. However, the reverse process is also possible—that women have lower social status as a result of gender differences in math achievement, attitudes, and affect. For example, the underrepresentation of women in research jobs might be, in part, a result of poorer performance in math by girls relative to boys; however, this cannot explain why girls perform more poorly than boys in some nations but not in others. The link between achievement and gender equity may be reciprocal, such that the underrepresentation of women in research jobs and the relatively poorer performance of girls in math in some nations perpetuate one another. Yet, it is much less plausible that indicators of gender equity such as enrollment ratios operate in this way. Although a dynamic and bidirectional process between societal gender equity and gender differences in math cannot be tested with these data, evidence of such a process would underscore the importance of initiatives aimed at improving both the social status of women and girls' math performance.

The finding that more negative attitudes and affect among girls (relative to boys) on the PISA were predicted by a greater women's share of parliamentary seats ran contrary to our predictions based on the gender stratification hypothesis. This finding appears contradictory when juxtaposed to the pattern of results in TIMSS math achievement and gender equity in enrollment ratios and research jobs. Yet, others have reported similar trends with regard to self-construals (Guimond et al., 2007), personality (Costa, Terracciano, & McCrae, 2001; McCrae, Terracciano, et al., 2005; Schmitt et al., 2008), emotion (Fischer & Manstead, 2000; Fischer, Rodriguez Mosquera, van Vianen, & Manstead, 2004), and values (Schwartz & Rubel, 2005). Most germane to the current study are the findings of Fischer and Manstead (2000), who analyzed gender differences in seven emotions across 37 cultures. They found larger gender differences in emotional intensity, duration, and experience in nations with higher GEM scores than in nations with lower GEM scores. This pattern was shown with the emotion of fear (as well as the six other emotions studied) and is thus relevant to our results on anxiety. How can we explain this apparently widespread finding that greater societal gender equity is associated with larger gender differences in some psychological constructs?

To explain why cultures that are more gender stratified actually tend to demonstrate patterns of gender similarities in many, though not all, psychological constructs, Guimond et al. (2007) integrated several social psychological theories about self-construal, social comparison, and power distance. Drawing from the work of Cross and Madson (1997), Guimond et al. proposed that gender differences in self-construals (a) account for gender differences in values, motivations, and emotions and (b) exist under conditions in which individuals make intergroup social comparisons (i.e., when they compare themselves to an out-group, such as the other gender) but not when individuals make intragroup social comparisons (i.e., when they compare themselves to their in-group, such as others within their gender). Consistent with theories of shifting standards in social comparisons (Biernat, 2005), individuals tend to self-stereotype or assimilate when they make intergroup social comparisons but tend to display contrast effects when they make intragroup social comparisons. Thus, girls may characterize their math anxiety as relatively high if their comparison group is boys but as relatively low if their comparison group is other girls. The tendency to engage in intergroup or intragroup social comparison varies cross-culturally, according to power distance (Guimond et al., 2007). Power distance is a cultural dimension that refers to the extent to which a society is stratified on the basis of social characteristics such as gender (Hofstede, 1980). High power-distance cultures (PDCs) tend to have rigid social hierarchies and to prohibit informal interactions between people of differing social strata, whereas lower PDCs tend to have more egalitarian social structures that encourage interactions among people of differing strata. As such, high PDCs discourage intergroup social comparisons in favor of intragroup social comparisons, whereas low PDCs do the reverse. High PDCs are inherently gender unequal, whereas low PDCs are inherently more gender equitable. Thus, individuals from high PDCs (e.g., Saudi Arabia) tend to engage in within-gender comparisons, resulting in a pattern of gender similarities. In contrast, individuals from low PDCs (e.g., Norway) tend to engage in between-gender comparisons, resulting in a pattern of gender differences. That gender equity was negatively associated with gender differences in math attitudes and affect in the current study is consistent with this theoretical perspective.

An alternative explanation of these unexpected findings draws on Maslow's theory of the hierarchy of needs (1954), which maintains that higher needs, such as feeling confident about one's abilities, require better economic, political, and educational conditions than do more basic needs, such as finding employment or caring for one's family. Individuals who are so oppressed and impoverished that they struggle to meet their survival and security needs cannot even think of meeting the esteem needs near the top of the pyramid. Girls who live in a society where they face major inequities because of their gender may be too consumed with meeting their most basic needs to think about such things as how anxious they feel about math or whether they feel good about their math skills. Consistent with this interpretation, nations with poor economic, educational, and political welfare generally tend to report the lowest mean levels (across both genders) of math anxiety, self-efficacy, self-confidence, and motivation (OECD, 2004). That is, the experience of math attitudes and affect may be a luxury, most often experienced by individuals who are not preoccupied with meeting more basic needs. This pattern of findings points to the importance not only of gender equity but also of human development.

Increases in women's agency—through improved education, economic opportunities, and property ownership—are central to human development (Sen, 1999; UNDP, 1995). Nobel laureate Amartya Sen maintained that

> the relative respect and regard for women's well-being is strongly influenced by such variables as women's ability to earn an independent income, to find employment outside the home, to have ownership rights and to have literacy and be educated participants in decisions within and outside the family. (Sen, 1999, p. 191)

That is, women's agency and empowerment are crucial to women's welfare. Thus, the assessment of women's individual and collective agency and empowerment is an important issue for study among social scientists (see special issue *7*(2) of the *Journal of Human Development* for reviews of this topic) and psychologists in particular (Grabe, in press). The varied convergent validity (as can be seen in Table 4) and predictive power of the four composite gender equity indices point to this issue. Although the indices tend to be composed of similar indicators—such as life expectancy, enrollment ratios, and parliamentary representation—the relative weight of each indicator influences the meaning and relevance of the indices. The GEQ is simple to compute, and data for most nations are readily available; however, it was not as reliably linked to the predicted gender gaps in math as the GEM and GGI were. Although the multidimensional nature of the GGI is appealing, the index is complicated and difficult to compute, even for the small number of nations with available data. Similarly, the GEM cannot be computed for all UNDP nations because so many of them lack high-quality estimates of its components. Nonetheless, from the perspective of the gender stratification hypothesis, the GEM was a good predictor of gender differences in math and a valid indicator of societal gender equity.

Future research on the inherently interdisciplinary topic of societal stratification by gender must address the complexities of measuring gender equity with composite indices. Specific domains of gender equity should be conceptualized as distinct from one another; gender equity in educational access and gender equity in political agency are not highly correlated and do not likely reflect a monolithic construct. Future research on societal gender equity must make precise predictions about the specific domain of gender equity that is theoretically relevant in order to be truly informative. The current study improves upon previous research in this regard.

Two international math assessments were meta-analyzed in the current study, and this provided an opportunity to compare findings between the data sets. When we compare the 18 nations participating in both TIMSS and PISA, it is apparent that the tests differ in their assessment of math achievement. The effect sizes found for gender differences in math achievement on TIMSS and PISA were not significantly correlated. TIMSS showed a stronger pattern of gender similarities, and PISA showed a pattern of a (albeit very small) male advantage in its content domains (notably, Space/Shape). In terms of attitudes, the magnitude of the gender difference on the PISA and TIMSS scales of Self-Concept and Self-Confidence in Math were highly correlated, though the PISA scale showed a significantly larger gender difference. Differences between TIMSS and PISA in these assessments have been noted by other researchers (Hutchison & Schagen, 2007) and remain an area of controversy among experts on cross-national math testing. These varying patterns of results in achievement may be understood best in the context of the explicit aims of TIMSS and PISA. Whereas TIMSS claims to assess the attained curriculum (i.e., what students have learned from the curriculum), PISA focuses on the concept of literacy, aiming to assess the abilities of students to use their mathematics knowledge and skills in the "real world." The TIMSS focus on the attained curriculum appears to be reflected in our finding that gender equity in enrollments significantly predicted the gender gap in math achievement on TIMSS but not PISA. Similarly, the pattern of gender differences found on PISA, with its focus on real-world applications, may be more sensitive to societal gender inequity. PISA may tap informal, out-of-school learning more than does TIMSS. Relative to girls, boys acquire more of this relevant knowledge from activities such as playing football, roaming about their neighborhoods, and playing videogames (Matthews, 1986, 1987). This might explain the slightly larger gender differences found on PISA than on TIMSS.

An alternative explanation for the larger gender differences on PISA is rooted in the greater male variability hypothesis, which would predict larger gender differences on more challenging math problems. Some have argued that PISA tests deeper mathematical skills that require students to integrate different existing kinds of knowledge (Hutchison & Schagen, 2007). Our analysis of the limited items made publicly available and the cognitive complexity classification schemes provided by the IEA and OECD supports such an argument. Three quarters of the items on TIMSS assess a DoK limited to recall and routine problem solving, with few items requiring students to use creative or strategic reasoning and problem-solving skills. In addition, whereas TIMSS tests recall of abstract knowledge of mathematics primarily with multiple-choice items, PISA tests application of mathematics primarily with constructed response items. Yet, our meta-analyses of the gender gap across the three cognitive domains on TIMSS indicate that boys did not perform better than girls on the more challenging problems that require creative or strategic reasoning and problem-solving skills. Although the results from this study demonstrate that TIMSS and PISA assess mathematics performance differently, both assessments showed very small gender differences at best. Thus, comparisons between TIMSS and PISA regarding test difficulty should not be overplayed as support for the greater male variability hypothesis.

Strengths and Limitations

None of the nations included in both PISA and TIMSS were characterized as low in human development in 2003 when students were tested, according to the UNDP. Nations low in human development have, by definition, poorer access to education; overall achievement on PISA and TIMSS assessments is strongly linked to human development. The nations with the lowest 2003 HDI values in the current analyses include Ghana, Morocco, Botswana, and Egypt. Of course, human development is much broader than what the HDI measures; by measures such as adult illiteracy rates and the probability at birth of not surviving to age 40, several of the nations in the TIMSS are conspicuously impoverished and underdeveloped. For example, Botswana ranks as one of the poorest nations in the world according to the UNDP's Human Poverty Index, and children in South Africa are less likely than children in most of the low HDI nations to reach Grade 5 in schooling (UNDP, 2003). It is likely that the limited access to formal schooling in less developed nations is greater for girls; nations with a medium or high HDI have significantly more equitable gender ratios in combined primary, secondary, and tertiary enrollment than do nations with a low HDI. Yet, high human development does not guarantee equal status of women: There are nations with high HDI values but very low GEM values (e.g., Japan and Korea). It is unclear how girls and boys in very low development nations compare in math achievement, attitudes, and affect or how societal gender stratification is reflected in those comparisons; if the gender stratification hypothesis is correct and our findings are not anomalous, we should expect larger effects with a broader sample. TIMSS and PISA—and secondary analyses of those assessments, such as the current study—would be strengthened if low human development nations were included in their samples. Of course, given that PISA includes only OECD nations, this limitation is more likely to be overcome by the IEA; indeed, TIMSS has generally increased its representation of less developed nations with each testing cycle.

Although studies in the 1990s demonstrated that gender differences in mathematics achievement emerge during adolescence (e.g., Hyde, Fennema, & Lamon, 1990), more recent data suggest that this gap has closed, at least in the United States (Hyde et al., 2008). The results of the current study are particularly significant in this developmental context because the students tested were between 14 and 16 years of age, precisely the ages when gender differences emerged in earlier studies. The overall pattern in the current data for many nations was one of gender similarities.

The current meta-analysis included math achievement, attitudes, and affect scores from 493,495 students across 69 nations. As a result of the statistical power afforded by the large TIMSS and PISA samples, the mean effect sizes reported here are reliable estimates of gender differences in math achievement, attitudes, and affect in the population. Yet, because reliable data necessary for the computation of both composite indices and domain-specific indicators of gender equity were not available for some nations, the multiple regression analyses used to test the gender stratification hypothesis had sufficient statistical power to detect medium but not small effect sizes ($R^2 \geq .10$; Cohen, 1988). This limited statistical power requires that we make conservative conclusions.

Conclusions

This meta-analysis provides further evidence that, on average, males and females differ very little in mathematics achievement, despite more positive math attitudes and affect among males. Yet, these findings of mean similarities in math are qualified by substantial variability across nations. Moderator analyses indicated that considerable cross-national variability in the gender gap can be explained by important national characteristics reflecting the status and welfare of women. Also, while differences within education systems, schools, and classrooms are critical influences on student achievement on TIMSS and PISA (OECD, 2004), the magnitude of gender differences in math also depends, in part, upon the quality of the assessment of mathematics achievement. If math assessments measure abilities that are not present in the curriculum, societal perceptions of their importance are likely to affect informal learning experiences and, in turn, achievement. Factors that have more direct influences on children's learning—for example, quality of instruction and curriculum—may serve to mediate the effect of gender inequity on math achievement. In particular, the value that schools, teachers, families, and children place on girls learning math should be a focus of future cross-national research on the gender stratification hypothesis. Consistent with the predictions of the gender stratification hypothesis and the psychological theories posited by Eccles (1994); Bandura (1986), and Eagly and Wood (1999), girls will perform at the same level as their male classmates when they are encouraged to succeed, are given the necessary educational tools, and have visible female role models excelling in mathematics.

References

Aronson, J., & McGlone, M. S. (2008). Stereotype and social identity threat. In Nelson, T. (Ed.), *The handbook of prejudice, stereotyping, and discrimination* (pp. 153–178). New York, NY: Psychology Press.

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education, 66,* 91–103.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development, 72,* 187–206.

Bhana, D. (2005). "I'm the best in maths. Boys rule, girls drool." Masculinities, mathematics, and primary schooling. *Perspectives in Education, 23,* 1–10.

Biernat, M. (2005). *Standards and expectancies: Contrast and assimilation in judgments of self and others.* New York, NY: Routledge.

Blascovich, J., Spencer, S. J., Quinn, D. M., & Steele, C. M. (2001). Stereotype threat and the cardiovascular reactivity of African-Americans. *Psychological Science, 12,* 225–229.

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review, 106,* 676–713.

Byrnes, J. P. (2005). Gender differences in math: Cognitive processes in an expanded framework. In A. Gallagher & J. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 73–98). New York, NY: Cambridge University Press.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology, 33,* 669–680.

Ceci, S. J., & Williams, W. M. (2007). *Why aren't more women in science?*

Top researchers debate the evidence. Washington, DC: American Psychological Association.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81,* 322–331.

Cross, S. E., & Madson, L. (1997). Models of the self: Self-construals and gender. *Psychological Bulletin, 122,* 5–37.

Dijkstra, A. G. (2002). Revisiting UNDP's GDI and GEM: Towards an alternative. *Social Indicators Research, 57,* 301–338.

Dijkstra, A. G. (2006). Towards a fresh start in measuring gender equality: A contribution to the debate. *Journal of Human Development, 7,* 275–283.

Dwyer, C. A., & Johnson, L. M. (1997). Grades, accomplishments and correlates. In W. A. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 127–156). Mahwah, NJ: Erlbaum.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Hillsdale, NH: Erlbaum.

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior. *American Psychologist, 54,* 408–423.

Eccles, J. S. (1994). Understanding women's educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *Psychology of Women Quarterly, 18,* 585–610.

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics, 21,* 55–69.

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization, and affective factors. *American Educational Research Journal, 14,* 51–71.

Fischer, A., & Manstead, A. (2000). The relation between gender and emotion in different cultures. In A. Fischer (Ed.), *Gender and emotion: Social psychological perspectives* (pp. 71–94). London, England: Cambridge University Press.

Fischer, A. H., Rodriguez Mosquera, P. M., van Vianen, A. E. M., & Manstead, A. S. R. (2004). Gender and culture differences in emotion. *Emotion, 4,* 87–94.

Fredricks, J. A., & Eccles, J. S. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38,* 519–533.

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology, 74,* 435–452.

Grabe, S. (in press). Women's human rights and empowerment in a transnational, globalized context: What's psychology got to do with it? In M. Paludi (Ed.), *Feminism and women's rights worldwide.* Westport, CT: Praeger Publishers/Greenwood Publishing Group.

Guimond, S., Branscombe, N. R., Brunot, S., Buunk, A. P., Chatard, A., Désert, M., . . . Yzerbyt, V. (2007). Culture, gender, and the self: Variations and impact of social comparison processes. *Journal of Personality and Social Psychology, 92,* 118–1134.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008, May 30). Culture, gender, and math. *Science, 320,* 1164–1165.

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

Hausmann, R., Tyson, L. D., & Zahidi, S. (2007). *The global gender gap report 2007.* Geneva, Switzerland: World Economic Forum.

Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 14–50). Baltimore, MD: Johns Hopkins University Press.

Hedges, L. V., & Nowell, A. (1995, July 7). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills, CA: Sage.

Hutchison, D., & Schagen, I. (2007). Comparisons between PISA and TIMSS: Are we the man with two watches? In Loveless, T. (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 227–261). Washington, DC: Brookings.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60,* 581–592.

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139–155.

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

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A., & Williams, C. (2008, July 25). Gender similarities characterize math performance. *Science, 321,* 494–495.

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

International Labour Organization. (1990). *International standard classification of occupations.* Geneva, Switzerland: Author.

International Association for the Evaluation of International Achievement. (2007). *TIMSS 2003 mathematics items: Released set eighth grade.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Jacobs, J., Davis-Kean, P., Bleeker, M., Eccles, J., & Malanchuk, O. (2005). "I can, but I don't want to": The impact of parents, interests, and activities on gender differences in math. In A. Gallagher & J. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 73–98). New York, NY: Cambridge University Press.

Kenney-Benson, G., Pomerantz, E., Ryan, A., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology, 42,* 11–26.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105,* 198–214.

Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research, 41,* 63–76.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56,* 1479–1498.

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

Machin, S., & Pekkarinen, T. (2008, November 28). Global sex differences in test score variability. *Science, 322,* 1331–1332.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Maslow, A. H. (1954). *Motivation and personality.* New York, NY: Harper & Row.

Matthews, M. H. (1986). The influence of gender on the environmental cognition of young boys and girls. *Journal of Genetic Psychology, 147,* 295–302.

Matthews, M. H. (1987). Sex differences in spatial competence: The ability of young children to map "primed" unfamiliar environments. *Educational Psychology, 7,* 77–90.

McCrae, R. R., Terracciano, A., & 79 members of the Personality Profiles of Cultures Project. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89,* 407–425.

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 of Research in Mathematics Education, 37,* 129–150.

Meece, J. L., Eccles-Parsons, J., Kaczala, C. M., Goff, S. B., & Futterman, R. (1982). Sex differences in math achievement: Toward a model of academic choice. *Psychological Bulletin, 91,* 324–348.

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its consequences for young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82,* 60–70.

Mullis, I. V., & Martin, M. O. (2007). TIMSS in perspective: Lessons learned from IEA's four decades of international mathematics assessments. In Loveless, T. (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 9–36). Washington, DC: Brookings.

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

Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains.* Chestnut Hill, MA: Boston College.

Mullis, I. V., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

National Academy of Sciences. (2006). *Beyond bias and barriers: Fulfilling the potential of women in academic science and engineering.* Washington, DC: National Academies Press.

National Research Council. (2006). *Learning to think spatially: GIS as a support system in the K–12 curriculum.* Washington, DC: National Academies Press.

OECD. (2003). *The PISA 2003 assessment framework.* Paris, France: Author.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris, France: Author.

Okagaki, L., & Frensch, P. (1994). Effects of videogame playing on measures of spatial performance: Gender effects in late adolescence. *Journal of Applied Developmental Psychology, 15,* 33–58.

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86,* 193–203.

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

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology, 76,* 114–128.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology, 94,* 396–404.

Porter, A. C., & Gamoran, A. (Eds.) (2002). *Methodological advances in cross-national surveys of educational achievement.* Washington, DC: National Academy Press.

Randel, B., Stevenson, H. W., & Witruk, E. (2000). Attitudes, beliefs, and mathematics achievement of German and Japanese high school students. *International Journal of Behavioral Development, 24,* 190–198.

Riegle-Crumb, C. (2005). The cross-national context of the gender gap in math and science. In L. Hedges & B. Schneider (Eds.), *The social organization of schooling* (pp. 227–243). New York, NY: Russell Sage Foundation.

Roberts, T.-A. (1991). Gender and the influence of evaluations on self-assessments in achievement settings. *Psychological Bulletin, 109,* 297–308.

Saccuzzo, D. P., Craig, A. S., Johnson, N. E., & Larson, G. E. (1996). Gender differences in dynamic spatial abilities. *Personality and Individual Differences, 21,* 599–607.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94,* 168–182.

Schüler, D. (2006). The uses and misuses of the Gender-Related Development Index and Gender Empowerment Measure: A review of the literature. *Journal of Human Development, 7,* 161–181.

Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: Cross-cultural and multimethod studies. *Journal of Personality and Social Psychology, 89,* 1010–1028.

Sen, A. (1999). *Development as freedom.* New York, NY: Knopf.

Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation, 14,* 87–100.

Sorby, S. A. (2001). A course in spatial visualization and its impact on the retention of female engineering students. *Journal of Women and Minorities in Science and Engineering, 7,* 153–172.

Sorby, S. A., & Baartmans, B. J. (2000). The development and assessment of a course for enhancing the 3-D spatial visualization skills of first-year engineering students. *Journal of Engineering Education, 89,* 301–307.

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? *American Psychologist, 60,* 950–958.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4–28.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

Stetsenko, A., Little, T. D., Gordeeva, T., Grasshof, M., & Oettingen, G. (2000). Gender effects in children's beliefs about school performance: A cross-cultural study. *Child Development, 71,* 517–527.

Terlecki, M. S., & Newcombe, N. S. (2005). How important is the digital divide? The relation of computer and videogame usage to gender differences in mental rotation ability. *Sex Roles, 53,* 433–441.

United Nations Development Programme. (1995). *Human development report 1995.* New York, NY: Oxford University Press.

United Nations Development Programme. (2003). *Human development report 2003.* New York, NY: Oxford University Press.

UNESCO Institute for Statistics. (November 2006). Women in science: Under-represented and under-measured. *UIS Bulletin on Science and Technology Statistics (3).*

Uttal, D. H., Hand, L. L., & Newcombe, N. S. (2009, April). *Malleability of spatial cognition: Results of a meta-analysis.* Paper presented at the biennial meeting of the Society for Research in Child Development, Denver, CO.

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,* 250–270.

White, H. (1997). *Patterns of gender discrimination: An examination of the UNDP's Gender Development Index.* The Hague, the Netherlands: Institute of Social Studies.

Wiseman, A. W. (2008). A culture of (in)equality?: A cross-national study of gender parity and gender segregation in national school systems. *Research in Comparative & International Education, 3,* 179–201.

Received October 9, 2008
Revision received August 16, 2009
Accepted October 6, 2009 ■