Sex differences in academic achievement are not related to political, economic, or social equality

Gijsbert Stoet a,⁎, David C. Geary b

a School of Education, University of Glasgow, 11 Eldon Street, G3 6NH, Glasgow, UK
b Department of Psychological Sciences, University of Missouri, Columbia, MO, USA

Abstract

The national differences in gender equality in economic and political participation have garnered considerable attention as an explanation of boys’ better achievement in mathematics in most countries, but the debate is far from resolved. Using data from four international assessments of the academic achievement of 1.5 million 15 year olds (Programme for International Student Assessment, PISA), we demonstrate that the relation between sex differences in PISA achievement and national measures of gender equality is not consistent across assessments, and several of the positive findings are confounded by outliers. Further, for overall achievement across reading, mathematics, and science literacy girls outperformed boys in 70% of participating countries, including many with considerable gaps in economic and political equality, and they fell behind in only 4% of countries. The results raise doubts about the relation between national equality policies and mathematics achievement, and raise broader questions regarding women’s underrepresentation in political, economic, and academic leadership despite stronger academic skills and regarding the long-term economic prospects and social stability of nations with many men who are not competitive in the modern economy.

Keywords:
Sex differences
Scholastic performance
Reading
Mathematics
Science

1. Introduction

The discussion of sex differences in educational attitudes and achievement dates back hundreds of years. For example, the British philosopher John Locke (1693) mentioned the relative ease with which girls are able to learn a second language in his treatise Some thoughts concerning education. The systematic study of sex differences in educational achievement started in the first half of the 20th century (Woolley, 1914). Overall, for the past 100 years girls have been found to perform better than boys in academic areas that involve language skills, including reading and spelling (Burt & Moore, 1912; Halpern, 2012), and boys in some mathematical areas, especially in adolescence and adulthood (Geary, 1996; Stockard & Bell, 1916). The stability and magnitude of these differences are vigorously debated, although a recent meta-analysis of sex differences in school grades from 1914–2011 concluded that the gap remained stable (Voyer & Voyer, 2014). This debate directly relates to a core question in the field of differential psychology and cognitive abilities, namely the degree to which sex differences in cognitive abilities are related to environmental factors. Some researchers have argued that these sex differences have been largely stable over the decades (Ellis et al., 2008), while others have argued they are disappearing (Feingold, 1988), or have essentially disappeared (Hyde, 2005; Hyde & Linn, 1988; Hyde & Mertz, 2009).

The purported disappearance of these sex differences, especially in mathematics, has been attributed to historical changes in social roles and movement toward gender parity in economic and political influence. This process is predicted by the gender similarities and gender stratification hypotheses, and some researchers have concluded that there is a linear relation between the size of the sex difference in mathematics
The concept of gender stratification has its origin in sociology (Collins, Chaletz, Blumberg, Coltrane, & Turner, 1993), referring to the differentiation between men and women based on income and power. Hyde's gender similarities hypothesis about gender differences in psychological variables posits that males and females “are more alike than different” (Hyde, 2005; Hyde, Lindberg, Linn, Ellis, & Williams, 2008). On the basis of a meta-analysis of some constructs, she concludes that there are only a few psychological variables on which men and women differ, such as some motor and sexual behaviors (Hyde, 2005). One of Hyde's main points is that the focus on sex differences results in an underestimation of girls' potential in mathematics.

The gist of the stratification hypothesis is that sex differences in economic and educational opportunities result in sex differences in mathematics performance (Else-Quest et al., 2010; Guiso et al., 2008). In short, this theoretical model proposes a causal link between gender equality and educational outcomes. The gender similarities and gender stratification hypotheses form a coherent theoretical model, which essentially states that nearly all psychological sex differences are the result of social and political factors. It thus follows that implementing policies that create equality of opportunity will ultimately eliminate the stratification of society by gender, and so eliminate any sex differences in achievement. It should be noted, though, that this theoretical model has been criticized for overlooking many psychological constructs that are quite large (e.g., occupational interests, Su, Rounds, & Armstrong, 2009), for creating the impression that psychological sex differences are categorical, that is, an all or none issue (Lippa, 2006), and for being based on inadequate methodology (Del Giudice, Booth, & Irwing, 2012). Nevertheless, the gender similarities hypothesis has received and continues to receive much attention in both academia and popular science and it is therefore important to subject the associated predictions to rigorous empirical tests. We specifically address the claim that indicators of social, political, and economic equality are related to the sex difference in mathematics achievement, and draw the attention of the field to a wider problem, that of boys' overall underachievement throughout most of the world.

The PISA is well suited for such an assessment, as it not only measures achievement in mathematics, reading, and scientific literacy in 15-year olds around the world, but also a considerable number of demographic and socioeconomic variables. The first PISA assessment was conducted in 2000, and repeated every three years since then, with an increasing number of countries, participating schools, and students across assessments; in the 2009 assessment, 515,958 students in 18,641 schools in 74 countries and economic regions participated (see Methods and Supplementary online material [SOM] for details).

Performance on the PISA is also linked to national differences in economic output, although it is unclear to what degree the national differences on the PISA (across the mathematics, reading, and science literacy scales) reflect general intelligence, variation in educational systems, or most likely some combination (Hunt & Wittmann, 2008; Lynn & Mikk, 2009; Rindermann, 2008). In any event, PISA scores matter; for example, it has been argued that an increase of 25 PISA points (in mathematics and science) in the next twenty years would raise the GDP by 115 trillion US dollars across OECD countries (OECD, 2006, p.27). In short, no other international instrument matches the PISA in terms of assessed breadth of educational achievement and related factors or in terms of assessed human capital and the economic well-being of countries and individuals. The one drawback is the narrow time frame of the PISA, which limits its sensitivity to long-term secular changes in achievement and factors that influence any such changes.

The results from previous comparisons of boys' and girls' performance in PISA assessments, and especially how differences in performance might be related to differences in gender equality measures (below) are confusing to say the least. Some studies, published in highly visible journals, conclude that there is a link between sex differences on the PISA and national gender equality policies (Else-Quest et al., 2010; Guiso et al., 2008; Hyde & Mertz, 2009; Reilly, 2012), whereas others do not find such a link (Fryer & Levitt, 2010; Kane & Mertz, 2012; Stoet & Geary, 2013).

Our approach to resolving the contradictory conclusions is to analyze all of the PISA assessments between 2000 and 2010. If the gender similarities and gender stratification hypotheses are correct, countries with higher levels of gender equality have smaller sex differences in educational achievement in mathematics as well as reading and science literacy. Previous tests of the hypotheses have focused primarily on mathematics and thus we do as well. Further, we argue that the focus is often on girls and mathematics, while boys' performance gets less attention, despite boys falling behind in reading around the world (Stoet & Geary, 2013).

We will show that when overall scholastic achievement is calculated from the core competencies in mathematics, reading, and scientific literacy, boys fell behind in the majority of countries. This is not only relevant for the gender stratification hypothesis, because it raises the question of why – despite educational opportunities and success – women are under-represented in leadership positions in politics, business, and academia. We will present support for two factors that might explain this. First of all, we show that the sex difference in educational achievement depends on the overall level of achievement, with boys at the highest levels doing equally well or better than girls at the highest levels. The second factor is relative academic strength. Students’ relative academic strengths and weaknesses are critical to understanding sex differences in later schooling and occupational outcomes (Humphreys, Lubinski, & Yao, 1993). Students who are relatively better at language related competencies than mathematics are more likely to choose humanities majors in college, whereas students who are relatively better at mathematics than language related competencies are more likely to choose physical science, technology, engineering, and mathematics majors (Park, Lubinski, & Benbow, 2007). These differences in turn contribute to later occupational choices and success within these occupations (Park et al., 2007). In other words, when it comes to choices made by individual students, mean sex differences in mathematics or reading may not be as important as whether they are, as individuals, better at mathematics than...
verbal competencies, including reading, or vice versa (Valla & Ceci, 2014; Wai, Lubinski, & Benbow, 2005; Wang, Eccles, & Kenny, 2013). We acknowledge there are other important factors influencing women’s underrepresentation in leadership positions that are beyond the current research project, especially factors that affect later career development, such as the pipe-line issues (Jacobs & Simpkins, 2005; Jeffcoat, 2004; Tucker, Hanuscin, & Bearnes, 2008). Further, there are a number of relevant sex differences in cognitive (Hedges & Nowell, 1995; Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995) and personality (Del Giudice et al., 2012; Irving, Booth, & Batey, 2014; Lippa, Collaer, & Peters, 2010; Schmidt, 2011; Su et al., 2009) traits (including interest differences) that influence sex differences in career choices and are beyond the scope of our data.

2. Methods

The current study combines the analysis of data from the PISA and equality data from a number of reliable international institutions, such as the United Nations. We will describe both types of data separately.

2.1. PISA data

The PISA data were collected in four international surveys carried out by the Organisation for Economic Co-operation and Development (OECD). The main aim of the surveys is to evaluate how well 15-year-old students (defined as between 15 years and 3 months and 16 years and 2 months) perform at the end of compulsory schooling (which in most countries is 15 years) in topics essential for employment and day-to-day functioning in modern economies, in particular reading, mathematics, and science (for an indepth descriptions of the PISA assessment framework, see OECD, 2003a, 2010). Data of the four surveys were collected in 2000, 2003, 2006, and the period 2009–2010. The PISA sampling and quality monitoring is documented in high detail; documents can be freely downloaded from the PISA website: http://www.oecd.org/pisa/

Since the first PISA survey in 2000, between 41 and 74 different countries and “economic regions” have participated (Table 1 lists countries for each PISA and OECD membership). The participating regions are parts of countries that did not participate as whole nations. For example, three economic regions in China participated: Shanghai, Hong Kong, and Macao. It should be noted that the total population of Shanghai (over 20 million) is larger than that of many participating whole nations, and some whole nations have fewer than one million inhabitants (e.g., Iceland). Similarly, in 2009 the two Indian states Himachal Pradesh and Tamil Nadu participated as economic regions while India as a whole did not participate.

In each participating country or region, a representative sample of schools and then a representative sample of participating students of each school are determined. The PISA consortium determines which schools participate. Schools then provide lists of students that are eligible to participate, from which national project managers select students according to standardized procedures (OECD, 2012, p.58). The OECD reports that on average between 4500 and 10,000 students per country are tested (OECD, 2003a).

From each survey, we have the data of each participating student. Those data include the standardized achievement scores on reading, mathematics, and science (and depending on survey year some additional basic skills, such as problem solving, or specific subtypes of mathematics, reading, or scientific literacy). From each student, we have also detailed information about background, including a number of socio-economic variables (including a standardized composite measure of socioeconomic status) and attitudes to learning. Altogether, we have hundreds of variables per participating student, including sample weighting variables.

Not every participating student completes all the test material. Instead, different students complete different test items such that each of them spends 2 h on the assessment.

The core PISA achievement (mathematics, reading, and science) levels are estimated using a Rasch model (which is commonly used in large educational surveys). For each of these three domains, we have five plausible values and 80 replicate weights. The values are based on the specific questions each student answered. The final scales for mathematics, reading, and science are transformed to have a mean value of 500 PISA points in OECD countries, with a standard deviation of 100 points. Each analysis of the core domains was carried out separately on each of the five plausible values and then averaged (exactly in accordance with the PISA guidelines for statistical analysis, OECD, 2003b). Country averages on the relevant values used in the analyses of this study are listed in the tables of the supplementary online material (SOM).

2.2. Gender equality data

The PISA does not include national gender equality data but such data are available from a number of other international organizations, such as the United Nations (United Nations Development Programme, 2011) and the World Economic Forum (Haussmann, Tyson, & Zahidi, 2010). We used two different types of national data reflecting gender equality from these sources. These included specific indicators of gender equality, such as the gender ratio in parliament or sex differences in longevity and income, as well composite indicators that combine a number of specific indicators. Among the latter, are the frequently used Global Gender Gap Index (GGI) published annually by the World Economic Forum in the Global Gender Gap Report (Haussmann et al., 2010), and the Gender Empowerment Measure (GEM), reported annually in the United Nations’ Human Development Report (United Nations Development Programme, 2011). The GGI is considered to be an improvement over older composite indicators such as the GEM, in part because it is more comprehensive, and in part due to the way it corrects for confounding factors, such as nations’ affluence (Zentner & Mitura, 2012).

The GGI measure was first introduced in 2006, and yet has been used in the analyses of the 2003 PISA data (Else-Quest et al., 2010; Guiso et al., 2008; Hyde & Mertz, 2009). Because the purpose of this study is, in part, to evaluate the conclusions drawn in those articles, we have also correlated GGI with PISA data before 2006. Although this is not ideal, we believe it is reasonable given the scores on these measure do not change much from year to year. For example, for countries participating
The use of composite gender equality indicators is relatively novel and not without limitations (Klasen, 2006). Despite this, the advantage of composite indicators over specific indicators is that they are the result of a careful and targeted approach to measuring gender equality, whereas the specific indicators are (necessarily) less reliable, and less likely to give a broad picture of national gender equality.

The authors of one of the studies in which a link between gender equality measures and the mathematics gender gap was found on the 2003 PISA stated: “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” (Else-Quest et al., 2010, p.122). However, the same authors argued that more specific indicators of gender equality, namely women’s share of research positions, parliamentary seats, and the female economic activity rates were also directly related to the gender gap in mathematics PISA1: “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.” Therefore, we not only tested GEM and GGI, but also these three specific variables for all four PISA surveys.

For women’s share in research positions, we combined the data from the United Nations Educational, Scientific and Cultural Organization (UNESCO) (available through http://stats.uis.unesco.org) and the OECD (available through http://stats.oecd.org/). We used both data sets, because they partially provide data for different countries. When the data were available in only one of the two sources, we used that information, and when information was available from both data sets, we averaged them. Further, these data were not always available for each year, and in those cases we used data from the closest year available. If a preceding and succeeding year was available we used the average.

The specific indicator of women’s economic activity was taken from the Human Development Reports for the years in the 2006 or 2009 PISA, the 2006 GGI and the 2009 GGI are strongly correlated, r(47) = .94, p < .01.

Further, it is important to note that even though GEM and GGI are different measures, there is a relatively strong linear correlation between them. For the countries that participated in the 2006 PISA for which a GEM and GGI score was available, the correlation was r(41) = .83, p < .001. For the countries that participated in 2009, the correlation was r(58) = .75, p < .001.

Table 1 (continued)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Singapore</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Slovenia</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>*</td>
</tr>
<tr>
<td>Tamil Nadu (India)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>*</td>
</tr>
<tr>
<td>Thailand</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>*</td>
</tr>
<tr>
<td>Trinidad and Tobago</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>*</td>
</tr>
<tr>
<td>Tunisia</td>
<td>–</td>
<td>*</td>
<td>–</td>
<td>*</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>*</td>
</tr>
<tr>
<td>Uruguay</td>
<td>–</td>
<td>*</td>
<td>–</td>
<td>*</td>
</tr>
</tbody>
</table>

1 Reilly (2012) makes similar claims for the 2009 PISA data, but his data analysis is confounded by the way the achievement gap in high achieving children is calculated; see SOM for a detailed analysis.
2000 (United Nations Development Programme, 2002), 2003 (United Nations Development Programme, 2005), 2005 (United Nations Development Programme, 2007) (the data of 2006 are not available), and 2009 (United Nations Development Programme, 2011). Similarly, the seats of women in parliament were taken from the Human Development Reports.

It should be pointed out that each Human Development Report states explicitly that comparisons of labor statistics across countries should be “made with caution”. Also, there is the more general question of whether “economically active” reflects increased opportunities and fairness in society. Arguably, economic activity can also mean that some women have to work for the mere survival of their family. Further, the measures of the percentage of women in research are not as complete as the composite indicators. As noted, we had to combine some data sources, but do not have information on the percentage of women researchers in some large countries, including Australia, Brazil, Canada, Russia and the United States. It is for these reasons that composite gender equality indices have been created, and we feel that they are therefore more suitable than these specific indices, although we report the results for both composite and specific indices.

### 2.3 Data Analysis

We use Pearson correlations to analyze relations between sex differences on the PISA and national equality scores (with an α criterion of 5%). In our analyses, we only include the data of whole countries, and not for the few economic regions (e.g., Hong Kong), because we do not know if the equality data available for whole countries are representative for the selected economic region — it seems reasonable to assume that especially the economic, social, and educational development of urban areas such as Hong Kong and Shanghai is different from that of other areas of China, in particular rural areas (and the same is true for other economic regions in PISA).

Further, for the calculation of a correlation between GGI and the mathematics gap, we use two different methods, following a previous study (Guiso et al., 2008). Here, correlations were based on data using children from families whose socioeconomic status was higher than the median of their country. They argued in their supplementary online material that there is the potential that boys and girls drop out of school at different rates, and that this is potentially more likely for children below the median on socioeconomic status. However, they also stated that their findings are robust when including all children. In any case, we have reported both values for the GGI and mathematics gap correlations. For outlier analysis in all correlations, we used Cook’s distance (Cook, 1977). We determined a data point as an outlier when it was higher than the commonly used cutoff point of $4/(n - k - 1)$.

### 3 Results

We start with an analysis of the consistency of the relation between gender equality measures on the one hand and the sex differences in mathematics achievement on the other. We then report on overall sex differences in achievement, and end with an analysis of how overall academic achievement levels and intra-individual differences in relative strengths in reading or mathematics can lead to insights into the observed sex differences in this study and in career choices.

The first part of our analysis challenges the conclusion that sex differences in mathematics achievement across countries are related to varying levels of gender equality. We start by showing that the correlation between the level of gender equality and the mathematics gap is not consistently found across the four PISA assessments (2000–2009). Next, we show that the observed correlations are very sensitive to the data of a few countries that can be considered as outliers.

The relation between the GGI and the mathematics gap is shown in Fig. 1. For each PISA assessment, we calculated the correlation twice; first using the method of Guiso et al. (2008), which removed school children from the lower half of each country’s socioeconomic status (see Methods), and second including all participating school children.

Of the six correlations, the mathematics gap and gender equality measures were only significant for the 2003 PISA, $r(35) = -0.41, p < .012$, irrespective of whether the data of all students are included or not (with all data: $r(35) = -0.44, p < .01$). Although the GGI measures were first reported for the year 2006, as noted above, we correlated the 2006 GGI with the 2000 PISA mathematics gap as well, and found no relation, $r(39) = -0.11, p = .49$. Thus, the prediction that higher levels of gender equality lead to a smaller sex differences in mathematics achievement was supported by the correlation between 2006 GGI and 2003 PISA data, but not supported for the 2000, 2006, or 2009 PISA data.

We carried out the same analyses using the main alternative of GGI, namely the GEM, which is available for all four years the PISA surveys were conducted (see Methods for explanation of differences between GEM and GGI). As with the GGI, there was a significant correlation between GEM and the mathematics gap in the 2003 PISA data set, with higher levels of gender equality associated with smaller sex differences in mathematics achievement, $r(31) = -0.367, p = .036$. However, this relation was not found for the 2000 PISA, $r(30) = .005, p = .977$, or the 2006 PISA, $r(41) = -0.027, p = .863$. In the 2009 PISA, we found a correlation in the opposite direction, $r(60) = -0.26, p = .043$: higher levels of gender equality were associated with larger sex differences in mathematics achievement. Thus again, the analysis shows that the prediction of greater gender equality leading to smaller sex differences in mathematics achievement was only supported for the 2003 PISA.

As explained in the Methods, there are three specific indicators that have been described as being related to the mathematics gap, and we tested these for all four PISA surveys as well. We found that women’s share of research positions correlated negatively with the sex differences in mathematics achievement, although not reliably so across all four PISA assessments (Fig. 2). Note that a negative correlation means that the larger the percentage of adult women in research positions, the smaller the sex difference in mathematics performance among school children. We observed this effect for the 2000 PISA, $r(31) = -0.68, p < .001$, 2003 PISA, $r(29) = -0.35, p = .058$, and 2006 PISA, $r(40) = -0.29, p = .07$ (trend), but not for 2009 PISA data, $r(46) = -0.16, p = .28$.

---

2 Else-Quest and colleagues used the data of 2001 from HDR report 2003 to correlate it with the 2003 PISA data (United Nations Development Programme, 2003).
The second specific indicator is female economic activity. Here, there was a clear effect for the 2000 PISA, $r'(39) = .352$, $p = .024$ and 2003 PISA $r(35) = .377$, $p = .021$. These positive correlations mean that the more adult women were participating in the economy, the smaller the sex difference among school children. However, this was not found for the 2006 PISA, $r(49) = -.046$, $p = .751$, or for the 2009 PISA, $r(62) = -.002$, $p = .986$. This latter finding is also true using just the countries that participated in 2000 and 2003 (for 2006: $r(40) = .282$, $p = .07$; for 2009: $r(42) = .089$, $p = .565$). Thus, the indicator female economic activity does not consistently relate to the mathematics gap, with a clear distinction between 2000–2003 PISA and 2006–2009 PISA.

The final specific indicator is the percentage of women in parliament. When applied to all available data, the correlation between this indicator and the mathematics gap never reached statistical significance, with correlations ranging between $r(37) = -.233$, $p = .154$ and $r(64) = .181$, $p = .146$. However, for the 2003 data, the expected negative correlation of a smaller mathematics gap with more women in parliament was found when the non-OECD countries were excluded, $r(28) = -.398$, $p = .030$ (Fig. 3).

Thus far, we have shown that the relation between composite and specific indicators and the mathematics gap was never found in more than two of the four PISA surveys, and even when found the pattern was not always the same. We now address one of the reasons why this relation was sometimes found and sometimes not.

A visual inspection of these correlational analyses suggests that grouping and outliers complicate the interpretation of statistically significant correlations. Before, we concluded that the GGI correlated with the mathematics gap in 2003 (but not other years). But even in 2003, this conclusion seems to be strongly influenced by the data from the Nordic countries, Iceland, Finland, Sweden, and Norway (Fig. 4). Iceland was a clear outlier in terms of Cook’s distance. For this dataset, a Cook’s distance value over 0.12 can be considered an outlier, and Iceland’s value of 0.38 was the only data point well over that value. If the same outlier analysis is applied to only half the PISA data, as in the paper by Guiso et al. (2008), then Iceland’s Cook’s distance is 0.62. The correlation was reduced to $r(34) = -.33$, $p = .053$, without Iceland, and to $r(31) = -.25$, $p = .16$ without the Nordic countries. Thus only a few data points might be the reason why the GGI was correlated with the mathematics gap in 2003 but not in any other PISA assessment. In the Discussion, we return to the contribution of the Nordic countries to the debate about sex differences in academic performance.

As is the case for the GGI data, the only correlation between GEM and the gap in mathematics was found for the 2003 data, $r(31) = -.367$, $p = .036$. Iceland, Thailand, and Korea had a Cook’s distance larger than the cutoff point, but even without these countries, the Pearson $r$ correlation still approached significance, $r(28) = -.34$, $p = .063$ (Fig. 5). It is interesting to note that the Nordic countries played an important role here as with the GGI. Dropping these countries and outliers reduced the correlation to $r(25) = -.25$, $p = .20$. In 2009, the positive correlation between the mathematics gap and GEM was entirely driven by the group of non-OECD countries. Again, we are not arguing that these countries should not be included, but we are merely demonstrating that some data points or groups of similar countries seem to have repeatedly played a key role in driving the correlation between gender equality measures and the PISA mathematics gap when it was found.

In regard to women’s share of research positions (Fig. 2), the 2003 results were strongly driven by four non-OECD countries. The OECD and non-OECD countries differed considerably in the percentage of female researchers, which particularly biased the data that Else-Quest et al. (2010) reported. The four non-OECD countries (Latvia, Thailand, Tunisia, Serbia) responsible for the highest percentages of female researchers (ranging from 47% to 53%) had a strong influence on the correlation. When excluding these four countries, the correlation was reduced from $r(29) = -.35$, $p = .06$ to $r(25) = -.172$, $p = .39$. Thus, the two PISA surveys for which there was a reliable relation between the mathematics gap and the percentage of women researchers irrespective of OECD status are 2000 and 2006. That in itself does not allow us to reject the hypothesis of a link between this variable and the sex differences in mathematics. However, using Cook’s distance, removing the outliers Iceland and Korea reduced the correlation between women’s share of research positions in 2003 and the mathematics gap to $r(27) = -.25$, $p = .20$.

In regard to the relation between the percentage of women in parliament and the mathematics gap, the only meaningful correlation was for the 2003 PISA when considering the OECD countries only, $r(28) = -.40$, $p = .03$. Removing the outlier Iceland from these data did not change the correlation.

In regard to the third specific gender equality indicator, the female economic activity rate, there was no indication that outliers affect the statistically significant correlations found for 2000 and 2003.

Altogether, the relation between gender equality indicators was unreliable across time, and when the relation existed, it was repeatedly driven by some countries, in particular Iceland and the nearby other Nordic countries.

### 3.1. The overall gap

The focus on international variation in gender equality and sex differences in achievement has largely been on the mathematics gap. This gap is certainly important, but so is the gap in “overall” achievement, that is the average achievement in the main PISA domains of mathematics, reading, and science. This is especially true given the strong correlations among these domains (in the 2009 PISA assessment, average $r$ for mathematics and reading = .81, for mathematics and science = .85, and for reading and science = .86, see SOM for further details about these cross correlations), and the importance of all three domains for success in employment and further education.

We found that boys’ overall score was lower than that of girls in most countries (Fig. 6). In the 2009 PISA, girls performed significantly better than boys in 70% of the countries (52 out of the 74), whereas boys significantly outperformed girls in only 4% of the countries (3 out of 74). In 2003 and 2006, the pattern was similar (in 2003, boys fell behind in 41% of the countries, girls in none; in 2006, boys fell behind in 58% of the countries and girls in 2%; the 2000 PISA data were organized differently such that inferential statistics for overall scores could not be calculated; see SOM for further details).

As with the mathematics gap, the overall gap was not consistently related to relevant national equality indicators. For
the 2003 PISA data a significant relation was found between GGI and the overall gap, \( r(35) = .403, p = .013 \), with boys falling behind more in gender equal societies. But also here Iceland’s Cook’s distance value was unusually high, and without Iceland the correlation was \( r(34) = -.31, p = .06 \).

Perhaps more important than between-country gaps are the overall achievement gaps found within each country. We calculated these by dividing the scores of each country into 100 percentiles, and then taking the difference (boys–girls). These scores, averaged for OECD and non-OECD countries revealed a

---

**Fig. 1.** Correlations between the composite gender equality index GGI on the one hand and the sex differences in mathematics achievement in PISA (boy scores–girl scores). Top panels) The relation between the 2006 GGI and the 2003 PISA mathematics gap. Note that the 2006 GGI data set is taken because 2006 is the first year GGI was published. The blue square indicates Iceland, and the three blue data points the three Nordic countries Norway, Sweden, and Finland. The left panel uses the data from students that score higher than the median socioeconomic status of a given countries (like Guiso et al., 2008), whereas the right panel data uses the data from all students. Middle panels) Same as top panels for the 2006 PISA gender gap in mathematics. Bottom panels) Same as top panels for the correlation between the 2009 GGI and 2009 PISA data.
quite remarkable pattern [Fig. 7]. The achievement gap, favoring girls, was larger among lower achieving children for both OECD and non-OECD countries. The only exception to the pattern of girls outperforming boys in overall achievement was found among the highest performers in OECD countries.

3.2. Achievement continuum and intra-individual differences

PISA data are well suited to assess intra-individual differences in reading and mathematics achievement, because scores are standardized; as noted, the average score in OECD countries is 500 ($SD = 100$) for mathematics and reading. We calculated this intra-individual gap by subtracting the mathematics and reading scores of each student (see SOM). Thus, if the value is positive, the student was better in mathematics, and if the value is negative the student was better in reading. When this mathematics-minus-reading gap is plotted by country (Fig. 8), a clear pattern emerges. In most countries, boys were better in mathematics than in reading, whereas girls were better in reading than in mathematics (that is, there was a common sex difference in the asymmetry of cognitive profiles).

4. Discussion

Our comprehensive analyses of all four PISA assessments enabled us to resolve current debate regarding whether or not a nation’s gender equality policies, customs, and women’s wider participation in politics and the economy is systematically related to sex differences in mathematics achievement and overall achievement, as predicted by the theoretical framework of the gender similarities and gender stratification hypotheses (Else-Quest et al., 2010; Guiso et al., 2008; Hyde & Mertz, 2009; Reilly, 2012).

Our main conclusions are as follows. First of all, there was no reliable relation between gender equality measures and sex differences in mathematics achievement or in overall achievement. It seems that the positive findings for the 2003 PISA data were unusual in that a small number of countries strongly contributed to the correlation; below we will discuss whether there is something different about those countries. Altogether, the results across the four very large data sets analyzed here are not consistent with the gender similarities and gender stratification hypotheses.

Our second conclusion is that there is a surprisingly large number of countries, including highly developed countries with progressive gender equality policies, in which boys fell considerably behind girls in overall achievement. In the 2009 PISA, this was true for 70% of the countries. Even though boys fell behind on average, this pattern was different among the top performers, especially in OECD countries. Here, top performing boys had higher achievement levels than did top performing girls. The achievement gap, favoring boys, at the high end may...
be one of the factors contributing to the overrepresentation of men in many leadership roles (e.g., academia, business), at least in OECD countries (of course, other factors play a role as well, as noted in the introduction). Our final conclusion is that individual boys, on average, have better mathematics skills than reading skills, whereas individual girls have, on average, better reading skills than mathematics skills, as noted by Valla and Ceci (2014), Wang et al. (2013) and Wai et al. (2005). We believe that this is one of the factors that contribute to the sex differences in college majors and careers. These conclusions all contradict with the gender similarities and stratification hypotheses.

4.1. Relation between measures of gender equality and sex differences in achievement are tenuous at best and likely spurious

Studies that have reported a relation between measures of a country’s gender equality and sex differences in mathematics achievement have largely focused on the 2003 PISA and concluded that countries with more equality have a smaller gap (Else-Quest et al., 2010; Guiso et al., 2008; Hyde & Mertz, 2009; Reilly, 2012). We also found such a relation for the 2003 PISA, but demonstrated that this was not consistently found in the other PISA surveys, and further that some countries seemed to drive the correlations (more about this below). This was the
case for composite gender equality measures, including the widely used GGI, and for specific measures (i.e., women’s participation in research, economic activity, and political participation).

Another issue highlighted by our re-analysis of previous studies is that there are many potential variables that can be correlated across the four different PISA assessments, and beyond that, there are multiple ways of correlating variables with the equality data of different years (e.g., in the reviewed studies, both the 2006 and 2007 GGI data have been correlated with the 2003 PISA). The large number of possible comparisons increases the risk of statistical type-I errors (Simmons, Nelson, & Simonsohn, 2011), that is, wrongly concluding that there is a relation when there is not one. In this view, strong arguments that national gaps in boys’ and girls’ achievement are related to national gender equality policies and outcomes are only supported if the relation is found consistently in each PISA assessment. And, they are not.

4.2. Are Nordic countries doing things differently?

Nordic countries (Iceland, Norway, Sweden, Finland, and Denmark) score very high in national gender equality measures (Hausmann et al., 2010), and there is much literature about the Nordic model in terms of equality policies. We have shown that, apart from Denmark, they play an important role in the correlations between equality measures and sex differences in achievement, especially in 2003. Therefore, one might argue that it is important to consider the approach of these countries to gender equality and education, and how their approach might be a model for other countries. However, we think that the situation is more complicated, as our data have shown. Finland, Iceland, and Norway are the top three OECD countries in terms of the overall achievement gap (whereas Sweden and Denmark are more similar to the other OECD countries, see SOM). Thus, even though these countries have a strong reputation for gender equality social policies, these policies clearly do not automatically lead to a reduction of sex differences in achievement. Alternatively, would there be reasons to argue that the Nordic equality model is the cause of boy’s poor overall achievement relative to girls? Is it fair to conclude that boys underperform at the cost of girls? We think that this conclusion is as unwarranted as the conclusion that gender equality leads to a reduction of educational achievement gaps. We believe that there is simply no systematic, causal relation. Our argument is further strengthened by our findings for Muslim countries. These countries generally have a poor gender equality record (Müller, 2006), and the patriarchally organized culture and Sharia inspired legal proceedings limit women’s

---

Fig. 5. The relation between the mathematics gap and the Gender Empowerment Measure (GEM). Lines indicate regression lines (red line only of OECD data points). The four panels show the data for the four PISA assessments and the GEM data of the corresponding year. Red circles represent OECD countries, and black circles of non-OECD countries. In 2009, there was a statistically significant correlation in the opposite direction, suggesting that the mathematics gap was larger in countries with a higher level of gender equality (but this was not the case for either the OECD or non-OECD countries taken apart).
and girls’ opportunities. Yet, in many of these countries, girls’ overall achievement is better than that of boys (see also Fryer & Levitt, 2010; Kane & Mertz, 2012).

There are important differences between the individual Muslim countries, as there are between the Nordic countries. We believe that these observations in regard to their educational achievement are most likely coincidences or at least not related to overall gender-related policies or customs. Instead, the finding that girls’ overall achievement is better than that of boys in countries as varied as Finland and Qatar demonstrates that political, economic, and social gender equality is largely uninformative about the sex differences in achievement.

At the other end of the spectrum are the countries in which boys perform statistically better than girls (namely Costa Rica, the Indian state Himachal Pradesh, and Columbia). The number of these countries is too small to look for similarities that can explain this, although you can always find something. For example they are all coffee-bean producing nations, but that does not justify the conclusion that coffee-bean production helps boys’ achievement; the same analogy-mistake might be true for the equality policies of Scandinavia and their girls’ relatively high achievement.

4.3. Girls outperform boys in most countries

As noted, most previous studies of sex differences in achievement have focused on mathematics. One result, in our opinion, is a comparative neglect of the gap in overall achievement (across mathematics, reading, and science), in particular the relatively poor achievement of boys in many parts of the world. Issues related to the overall achievement
gap have been examined previously (Gorard, Rees, & Salisbury, 2001; Hodgetts, 2008; Machin & McNally, 2005), but surprisingly (given the wealth of data) have received little attention in the context of PISA data. In every PISA assessment, the same patterns emerged: girls had higher overall achievement than boys, and the magnitude of the gap varied systematically across the performance continuum. The gap was the largest at the lowest levels of achievement and was closed in most developing countries and reversed in developed countries at the highest levels of achievement. Debates regarding sex differences in mathematics achievement and in STEM fields were sparked in these latter countries and for the highest achieving students.

We suggest that the focus on the mathematics gap in these countries has contributed to the lack of attention to boy’s academic underperformance overall and especially for reading. The finding that girls’ overall achievement exceeds that of boys’ in most countries raises a question that has not yet been adequately addressed: given that girls do so well academically, why is it that women are underrepresented in leadership positions in politics, business, and academia? This is a fundamentally important question, because answering it will not only help to better understand the relation between cognitive performance measures and real-world career outcomes, but it will also have major implications for the effectiveness of equality policies. There is a direct link between this question and the lack of a link between gender equality and educational outcomes. For example, despite the lack of opportunities for girls and women in Muslim countries, girls’ achievement is better than that of boys in those countries. This implies that opportunities for girls and educational outcome are unrelated (or if they are related it is not in a transparent way), and if they are unrelated, it is not surprising that their educational success does not directly translate in full socio-political and academic participation.

One often proposed answer is that women’s underrepresentation is the result of explicit or implicit sexism, although it is not the only factor (Ceci & Williams, 2011; Ceci, Williams, & Barnett, 2009). On the basis of our results, we also have to consider that the overall achievement gap between boys and girls is different for low and high performing children. The gap is large and favors girls at the lower end of performance, but it is non-existent or favors boys at the high end of performance (Fig. 7). Given that there is indeed evidence that a large proportion of the best performing school children end up in successful careers, some leaders in politics, business, and academia (Kell, Lubinski, & Benbow, 2013; Wai, 2013, in press), we might therefore expect an overrepresentation of boys just based on performance data alone, especially in OECD nations. We call this the “high-achievers male advantage gap” hypothesis. We believe that this is only one factor among many of the reasons for the leadership gap and clearly is not a viable factor in non-OECD nations. There are, however, other factors at play as well, and some of these start to come into play at later stages.

Fig. 8. Intra-individual gap between reading and mathematics. In each PISA assessment, boys were, on average, better in mathematics than in reading (x-axis), while girls were, on average, better in reading than in mathematics (y-axis). Each data point represents the data of one country.
of career development, such as pipe-line issues and other cognitive or personality variables mentioned in the introduction. At the very least, our study helps to understand the role of sex differences at the end of compulsory education. That helps to constrain possible models of what causes the underrepresentation of women in politics and some occupations. One of the interesting outstanding questions for future research is whether the increased opportunities for talent development in richer countries explain why this gap is only found in the group of OECD nations (Fig. 7). In OECD nations additional factors likely include the different trade-offs that women and men make with respect to investment in occupational attainment and their families (Ferriman, Lubinski, & Benbow, 2009), differences in career choices (Wang et al., 2013), and in competitive striving (Geary, 2010, pp.213–235; Sax & Harper, 2007).

4.4. Sex differences vary depending on where in the performance distribution school children are

One of the findings of our study is that while, on average, boys are outperformed academically by girls, the situation is different at the extremes of the distribution. As we reported previously for the PISA data and as is known from numerous other studies (Benbow & Stanley, 1982; Benbow & Stanley, 1983; Stanley & Benbow, 1982; Stoet & Geary, 2013; Wai, Lubinski, Benbow, & Steiger, 2010), the sex difference in mathematics performance is considerably larger at the high end of performance distributions. Interestingly, at least for PISA data, we found no sex differences at the low end of mathematics performance (Stoet & Geary, 2013). In regard to reading differences, the pattern is exactly the opposite, with the largest gap at the low end and the smallest gap at the high end (Stoet & Geary, 2013). And indeed, in the current study we found that while boys fell behind across subjects, it is the opposite for the children at the top, where boys are in the lead (Fig. 7).

4.5. Relative intra-individual strengths play role

Finally, studies of predictors of college majors and career trajectories have revealed the importance of intra-individual differences, that is, students’ relative academic strengths and weaknesses (Humphreys et al., 1993; Park et al., 2007), an issue that has not been assessed in previous studies of sex differences using the PISA. For instance, it is not simply the absolute level of mathematical competence that predicts entry into STEM fields, but also the level of competence relative to other academic skills. Individuals with strong mathematical abilities are more likely to pursue non-STEM careers if they also have strong verbal abilities, which include more women than men (Lubinski & Benbow, 2006). Our findings across all four PISA assessments revealed that in most countries boys have relatively better mathematics achievement and girls relatively better reading achievement, with implications for college and later occupational patterns, in particular high achievers entering STEM professions. These intra-individual differences may in fact be relatively more important than mean sex differences in pursuit of STEM careers, consistent with previous studies (Humphreys et al., 1993; Park et al., 2007; Valla & Ceci, 2014; Wai et al., 2005; Wang et al., 2013) and current sex differences in university enrollment (men are more strongly represented in STEM, and women more strongly represented in languages), especially when combined with large sex differences in vocational interests (Lubinski, 2000; Su et al., 2009). Further, these intra-individual differences do not vary across the achievement spectrum as do the general mathematics and reading gaps (Stoet & Geary, 2013); that is, in most countries, it is the case that boys are better in mathematics than in reading while girls are better in reading than in mathematics, irrespective of whether they are middle or high attainers (see SOM).

4.6. Outlook

We found no evidence for a reliable relation between PISA performance and gender equality. This not only casts doubt on the gender similarities and gender stratification hypotheses, but also has important implications for educational policy. Implications for educational policy go well beyond the data and require taking value frameworks into account. The current paper is not the right place to elaborate on such frameworks. Nevertheless, it is possible that the current data will influence how policy makers think about the options available, for example, to increase levels of equal opportunities in education. Therefore, we think it is important to raise a few issues that go beyond the interpretation of the data. We believe that one important implication of our analyses is that policy makers and educators should not expect that broad progress in social equality (e.g., increased political participation of women) will necessarily result in educational equality. In fact, we found that with the exception of high achievers, boys have poorer educational outcomes than girls around the world, independent of social equality indicators; a gap that has grown for reading over the past decade (Stoet & Geary, 2013). This has strong implications for national economic prospects as well as social stability, as young men who are not well integrated into the economic structures of their society are prone to crime and violence (Wilson & Daly, 1985). It also seems that the majority of research papers and interventions have focused on reducing the mathematics gap. Our finding that the mathematics gap has been stable for at least the first decade of the 21st century raises questions regarding the efficacy of the interventions that have been used thus far (Stoet & Geary, 2012), whether the interventions focus on the right segment of girls (closing the overall gap requires interventions with the highest achieving girls in mathematics Stoet & Geary, 2013), or whether the amount of resources devoted to these interventions has been insufficient. We certainly believe that increasing the numbers of women entering STEM professions and better understanding why this has been difficult to achieve are important goals (Ceci & Williams, 2011; Ceci et al., 2009), but the focus on this issue has resulted in a relative neglect of boys’ academic underperformance. This underperformance has been debated and solutions have been proposed (e.g., Hoff Sommers, 2001; Weaver-Hightower, 2003; Yates, 1997; Younger & Warrington, 2004), but our impression of the literature is that this stands in no comparison to the attention the underachievement and underrepresentation of girls in STEM subjects has received. In any case, psychological research focused on better understanding the reasons for this phenomenon and addressing them have the potential to improve the well-being of many people throughout much of the world.
Acknowledgments

We thank Lynne Cooper, Ben Winegard and Drew Bailey for insightful comments and discussion.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.intell.2014.11.006.

References


