In press in Intelligence. Accepted for publication on 6th April 2020.
Cite as: Stoet, G. & Geary, D.C. (in press). Sex-specific academic ability and attitude patterns in students across developed countries. Intelligence.

Sex-specific academic ability and attitude patterns in students across developed countries

Gijsbert Stoet, University of Essex, Colchester, UK
David C. Geary, University of Missouri, Columbia, Missouri

Abstract

The extent of sex differences in psychological traits is vigorously debated. We show that the overall sex difference in the pattern of adolescents' achievement and academic attitudes is relatively large and similar across countries. We used a binomial regression modeling approach to predict the sex of 15 and 16 year olds based on sets of academic ability and attitude variables in three cycles of the Programme for International Student Assessment (PISA) data (*N*=969,673 across 55 to 71 countries and regions). We found that the sex of students in any country can be reliably predicted based on regression models created from the data of all other countries, indicating a common (universal) sex-specific component. Averaged over three different PISA cycles (2009, 2012, 2015), the sex of 69% of students can be correctly classified using this approach, corresponding to a large effect. Moreover, the universal component of these sex differences is stronger in countries with relative income equality and women's participation in the labor force and politics. We conclude that patterns in academic sex differences are larger than hitherto thought and appear to become stronger when societies have more socioeconomic equality. We explore reasons why this may be the case and possible implications.

Sex differences in numerous personality and cognitive traits are well established (for reviews, see Archer, 2019; Geary, in press; Halpern, 2011; Lippa, 2005; Miller & Halpern, 2014), but there remains an ongoing debate regarding their magnitude. For example, Hyde (2005) argued that most of these sex differences are close-to-zero or small, whereas others argued that many of them are more substantial (Archer, 2019; Del Giudice, 2009; Del Giudice et al., 2012); even relatively small sex differences in educational variables can have large-scale consequences (Gibb et al., 2008).

One of the reasons for this lack of agreement resides in how traits are selected for inclusion in the relevant study or review. For example, Archer's (2019, Table 3) listing of sex differences is not only more detailed than Hyde's (2005, Table 1), his analysis is far more theory driven (i.e., including traits that have been under different types of evolutionary selection pressure). A second reason relates to the way sex differences are conceived. For example, both Hyde's (2005) and Archer's review of sex differences report the effect sizes of individual traits, whereas others report multivariate effect sizes (e.g., the pattern in personality traits; Del Giudice, 2009; Del Giudice et al., 2012). Note that throughout this article, we use the term "multivariate" to indicate that sex differences are calculated across multiple variables (see Methods for the how these calculations are performed).

These issues extend to sex differences in academic abilities, such as reading and mathematics achievement, and associated attitudes (e.g., mathematics self-efficacy), but at this point have only been assessed as single variables (i.e., one variable at the time rather than as sex differences in combinations of

multiple variables). We contribute to this field by examining the cross-national pattern of sex differences in academic abilities (reading, science, mathematics) and related attitudes using large international data sets (N = 969,673). Our study uses a novel method of determining how well children's sex can be predicted based on multivariate data patterns observed in other countries.

There is considerable international variation in the magnitude of the sex differences on individual measures of mathematics, reading, and science achievement (OECD, 2016). At the same time, there are complex and surprising relations between these sex differences. For example, the smaller the national sex difference in mathematics achievement, the larger the sex difference in reading achievement (Stoet & Geary, 2013). The trade-off between mathematics and reading achievement means that a single trait cannot capture the pattern of sex differences in academic abilities and will lead to an incomplete and potentially inaccurate assessment of the factors that contribute to them. For instance, the finding that the sex difference in mathematics achievement varies across countries and is negligible in some of them has been presented as evidence that any sex differences in mathematics are largely or solely caused by social and cultural factors (e.g., Else-Quest et al., 2010; Spelke, 2005).

While we do not doubt that culture can influence the expression of sex differences, we theorize that biological constraints make it difficult to completely eliminate them, as appears to be the case for many non-academic domains (e.g., aggression, sex-typical play patterns; Geary, in press). For instance, boys and men typically outperform girls and women on mathematical word problems because they tend to diagram the relations described in the problems and this in turn reduces errors (Geary et al., 2000; Lewis, 1989). The sex difference in the use of spatial strategies in this context stems from a broader and likely evolved male advantage in spatial abilities that are co-opted for academic learning (Geary, 1996). Interventions that teach girls and women to use diagrams reduce the sex difference on word problems but do not change the underlying differences in spatial abilities (Johnson, 1984). This means that the sex difference will remain when this intervention is not applied, or when it cannot be applied to the particular problem. Analogously, girls and women have likely evolved advantages in language abilities that contribute to their well-documented advantages in reading comprehension (Asperholm et al., 2019; Reilly et al., 2019). Interventions focused on boys' learning to read (e.g., phonemic awareness and word decoding) should reduce these gaps, but will not change underlying sex differences in language proclivity (e.g., ease of word learning and discrimination of basic language sounds; Majeres, 2007). Thus, sex differences in academic abilities might be reduced with sex-specific interventions, which will likely require boys and girls to spend different amounts of time on learning skills for which they are relatively weaker. Given that there is still limited support for such sex-specific interventions, certain sex differences will continue to be clearly expressed across the world.

In summary, we make two specific points. The first is that the magnitude of mean sex differences in academic outcomes might fluctuate across contexts but any such fluctuations are not independent of sex differences in other academic outcomes, necessitating the examination of patterns of abilities and attitudes. Second, we hypothesize that there are biologically influenced sex differences in cognitive abilities and interests that will result in consistent sex differences in academic and achievement profiles throughout the world, even when mean differences in one area or another fluctuate (Geary, in press, 1996).

The sex differences in academic abilities and attitudes are not only of theoretical interest, but also of a sociopolitical concern because they influence the occupational and educational choices of women and men (Stoet & Geary, 2015). For example, the finding that many adolescent girls fall behind in generic mathematics tests and are underrepresented in many science, technology, engineering, and mathematics (STEM) areas has led to the development of programs to support girls in these areas (Hag, 2002; Wang &

Degol, 2017). Similarly, the finding that boys' reading achievement falls behind that of girls has led to several policy-based programs to address this gap (e.g., Ontario Ministry of Education, 2004).

To examine sex-specific patterns in educational measures, we used data from three successive waves of the OECD Programme for International Student Assessment (PISA), which is the largest educational assessment data set in the world. The assessments included achievement in a variety of academic domains (e.g., mathematics) and attitudes (e.g., joy in reading). We hypothesized that there are consistent sex-specific academic and attitude patterns across countries. Our approach is to use logistic regression models to determine how well student sex can be predicted based on a number of different educational measures, as well as how well the predictive model of one country can be used to predict student sex in other countries.

This approach was inspired by attempts to predict sex from the pattern of gray and white matter in the brain (Chekroud et al., 2016; Del Giudice et al., 2016; Rosenblatt, 2016). The basic idea is that the pattern of gray and white matter (as identified with MRI scans) should be considered as a whole. Using logistic regression, the pattern of white and gray matter of one set of participants is used to create a classification model which can then be used to predict sex based on the pattern of brain data of other participants (which makes sense, given the well-documented sex differences in the brain, Bramble et al., 2017; Dean et al., 2018; Escorial et al., 2015; Jahanshad & Thompson, 2017; van der Linden et al., 2017). Using this approach, Checkroud et al. (2016) could predict sex with 93% accuracy. An analogy is provided by people's holistic processing of facial features to determine others' sex. Even though sex differences are smaller for individual facial features, the combinations allows people to quickly determine the sex of 19 out of 20 people (Bruce et al., 1993; Del Giudice, 2013). When it comes to sex differences, the whole is more than the sum of the parts.

Similarly, if there are systematic patterns across academic domains and attitudes, one should be able to classify student sex better than chance. If the patterns are universal (i.e., common across countries), a logistic regression model based on the data of one country should be able to predict students' sex in any other country.

Further, even if there are universal sex-specific patterns, the extent to which students in any one country fit this pattern might vary in systematic ways. In line with previous work on international variation in sex differences (Costa et al., 2001; Falk & Hermle, 2018; Schmitt et al., 2008; Stoet & Geary, 2018), we expected that any such differences would be larger in countries with higher levels of social, political, and economic equality. Although the exact reasons for such a correlation are still a matter of debate, it is important to determine if this rather paradoxical correlation – larger sex differences in more egalitarian countries – applies to the broad pattern of academic competencies and attitudes.

There are several reasons why such a correlation might emerge. The first is that sex differences in the underlying cognitive abilities generally become larger as general health and living conditions improve. The basic idea is that many traits that show sex differences have evolved to signal the health and resilience (e.g., to infection) of the individual and can only be fully developed in healthy individuals with low levels of exposure to disease, nutritional shortfalls, and other stressors (Cotton et al., 2004; Geary, 2015, 2016). Average sex differences in these traits, including spatial abilities (favoring men) and language abilities (favoring women), would then be larger in populations buffered from these stressors (see Geary, 2015). The larger sex differences in language and spatial abilities, as examples, would also manifest as larger sex differences in academic domains that are dependent on these abilities, as illustrated above. The second reason is that improvements in living conditions are often associated with the liberalization of educational policy, allowing students more freedom in their own academic choices (e.g.,

elective coursework) based on their interests and strengths. The sex differences in academic strengths and interests will be magnified by such choices (Stoet & Geary, 2015; Su et al., 2009).

Methods

PISA data

The PISA is an evaluation of academic achievement and attitudes that is conducted in three-yearly cycles by the Organisation for Economic and Cooperative Development (OECD) and partner countries. In each cycle, a representative sample of hundred thousands of students is administered a two hour assessment (OECD, 2012). Students sampled are between 15 years and three months and 16 years and two months of age at the time of assessment and should have completed at least six years of formal schooling. All test material is translated and, where necessary, specific concepts are adjusted to the local culture.

Students' competencies in the domains of reading comprehension, science literacy, and mathematics are assessed and their abilities in these areas are estimated using a sophisticated statistical model and results in numerical test scores for each participating student (for details see OECD, 2012). Each cycle includes achievement assessments in each of these three academic domains and in addition focuses on associated attitudes in one of the three domains (e.g., 2015 PISA focused on science; that is, most of the attitude variables were about science motivation and related behavior). We included three PISA cycles to capture the full range of academic domains and attitudes (2009 PISA: Reading; 2012 PISA: Mathematics; 2015 PISA: Science). Arguably, the results of one cycle are sufficient to prove the point, but demonstrating the same effect in multiple independent datasets would strengthen the conclusions, especially when different attitudes from quite different academic domains are used to classify sex.

The PISA not only samples students from separate countries, but also from a number of economically independent or semi-independent regions, such as Hong Kong. The PISA reports overlapping data for both country and region in some cases. We eliminated all such instances by excluding the separate datasets for the states of Florida, Massachusetts, and North Carolina in the United States, the Perm territory in Russia, and the regions dataset of Spain. We also excluded the data from Albania, because of PISA reported a mismatch between different test booklets, which makes identification of student sex unreliable (OECD, 2017, p. 269). We excluded Liechtenstein's data due to the unusually small sample size (in 2012, n=203) compared to a median sample size of over 5,245 students across the rest of the included datasets in the 2012 PISA cycle. A complete list of all included countries can be found in the supplementary online material.

All students in the PISA completed the tests in the domains of reading comprehension, science literacy, and mathematics, yet not all students completed the attitude surveys. We only included students for whom we had a full dataset (N = 969,673, see Table 1 for the numbers of participating boys and girls and countries).

PISA cycle	Cognitive variables used	Attitude variables used	Countries (*n*)	Girls (*n*)	Boys (*n*)
2009	Mathematics, Reading, Science	Enjoyment of reading, Library use, Online reading, Diversity of reading	71	247,763	236,180
2012	Mathematics, Reading, Science	Mathematics Self Concept, Interest in mathematics, Instrumental motivation for mathematics, Mathematics behavior, Mathematics anxiety, Mathematics self-efficacy, Mathematics intentions, Mathematics work ethic, Attributions to failure in mathematics, Subjective norms in mathematics	62	73,619	69,120
2015	Mathematics, Reading, Science	Interest in broad science topics, Science activities, Joy of science, Science self-efficacy, Instrumental motivation for science, Epistemological beliefs	55	176,091	166,900

Table 1: For each PISA cycle used in this study, we report the number of performance and attitude variables used in the logistic regressions, the number of included countries, and the total numbers of boys and girls with a complete data set.

Each student's scores on the reading comprehension, science literacy, and mathematics tests were available as three sets of plausible variables (five per domain in 2009 and 2012, and 10 per domain in the 2015 PISA cycle). In short, plausible values are often used in large-scale educational assessments, because each student is working with a different subset of test items from the total item pool, which makes it inappropriate to simply use the percentage of correctly solved items, because the different subsets of items might vary in difficulty (OECD, 2017, p. 128). Instead, PISA uses item response theory scaling. Plausible values are, in essence, random draws from possible values from a posteriori distribution for a given student. Working with plausible variables requires a special type of data analysis. That is, each analysis needs to be carried out with each different plausible variable set and resulting statistics are then averaged for the different plausible variable sets. PISA provides excellent documentation on how exactly to carry out such an analysis (as well as SPSS and SAS macros), which we have followed throughout (OECD, 2009).

The reading, science, and mathematics tests are not in the public domain, but the PISA documentation provides representative samples of items (OECD, 2018). For example, the 2015 PISA with a focus on science included three multiple choice questions about meteors and craters (which was one of the multiple units of the science questions). For these three questions, a context was provided to students in text and image format. The context was "Rocks in space that enter Earth's atmosphere are called meteoroids. Meteoroids heat up, and glow as they fall through Earth's atmosphere. Most meteoroids burn up before they hit Earth's surface. When a meteoroid hits Earth it can make a hole called a crater." The associated first question was "As a meteoroid approaches Earth and its atmosphere, it speeds up. Why does this happen?" with the following four possible answers (of which only one was correct): 1. The meteoroid is pulled in by the rotation of Earth. 2. The meteoroid is pushed by the light of the Sun. 3. The meteoroid is attracted to the mass of Earth. 4. The meteoroid is repelled by the vacuum of space. This question "required students to apply simple scientific knowledge to select the correct explanation for why objects

speed up at they approach Earth". The subsequent question about craters and meteors was "What is the effect of a planet's atmosphere on the number of craters on a planet's surface?". Students had to select "more/fewer" at two places in the following given sentence: "The thicker a planet's atmosphere is, the more/fewer craters its surface will have because more/fewer meteoroids will burn up in the atmosphere". This question required students "to select two responses that explain the relationship between the thickness of a planet's atmosphere, the likelihood that meteoroids will burn up in the atmosphere and, therefore, the number of craters that will be on the planet surface". Finally, students had to order three given craters by size and age based on a picture showing three different overlapping craters. This question "required simple, everyday knowledge that a larger object would cause a larger crater and a smaller one would cause a smaller crater" and it required students to "compare the three craters shown in the image to determine when the craters were formed, from oldest to newest, based on the way they overlap in the image."

In addition to the three achievement variables, we added the attitude variables related to the focus of the specific PISA cycle. For example, for the 2009 PISA, these variables were "diversity in reading material", "enjoyment of reading", "library use", and "online reading". The details of each measure are listed in Appendix A. There are more attitude variables not directly related to the domains mathematics, reading, and science, which we did therefore not include (e.g., the degree to which children enjoy cooperation or test anxiety).

The PISA is a complex instrument that reports details about the reliability of its scales in exhaustive technical reports (OECD, 2012, 2014, 2017). For example, for the 2015 PISA, the technical report lists the reliability of each scale for each country separately (OECD, 2017 p.232). Because the PISA uses "a rotated and incomplete assessment design" (OECD, 2017, p.231), it reports test reliability in terms of "explained variance" for each cognitive domain based on weighted posterior variance (which is the variance across the plausible values). The explained variance of the statistical model used reports values that range from .80 to .91 for the achievement and attitude scales (OECD, 2017, Table 12.4 and Table 12.5) suggesting that these are very reliable measures.

International indicators of income and women's empowerment

We used Global Gender Gap Index (GGGI) data from the Global Gender Gap Report for the years 2009, 2012, and 2015 (World Economic Forum, 2006, 2012, 2015). The GGGI reflects women's participation in the economy, in politics, as well as equality in "health and survival" and years of schooling. The GGGI score falls, in principle, on a scale between 0 (large gap) and 1.0 (no gap).

We used the Gini coefficient data from the World Bank for the years 2009, 2012, and 2015 available via https://data.worldbank.org/. The Gini coefficient reflects household income distribution, with potential values between 0 (maximum equality) and 100 (maximum inequality).

Analyses

We applied binomial logistic regressions (without interaction terms) to each national sample in each of the three PISA assessments to predict student sex. This has the advantage of being easy to implement while allowing for a multivariate weighting of the relative importance of one predictor relative to others in the model.

Note that there is a direct correspondence between a binomial logistic regression (without interaction terms) and the multivariate Mahalanobis distance (we will provide a specific example of this below). An

easy way to understand this concept is to imagine that each student in our study can be represented as a point in a multidimensional space (whereby each included variable is one of the dimensions). The Mahalanobis distance (also known as a multivariate Cohen's *d*) gives an indication of the distance between clusters of points (e.g., centered around boys' profile versus that of girls) in multidimensional space (for a review, see Del Giudice, 2013). The Mahalanobis distance is identical to the Cohen's *d* applied to the predicted values of the binomial logistic regression (without interaction terms).

For each country (and each PISA cycle), we then used the logistic regression model to predict student sex. These models (i.e., weightings for individual variables) were then combined with weights derived from other countries to predict the sex of students in all other countries. More precisely, based on the resulting matrix of weightings (i.e., regression coefficients for reading comprehension, mathematics, science literacy, and attitudes), we determined for each country the average percentage of students whose sex was successfully predicted based on the models derived from all other countries. For calculating the percentage of correctly predicted values, we used the student weights (each student in the data set has an assigned weight to compensate for varying sampling probabilities, OECD, 2009, p. 48). The sum of all student weights in a country equals the total number of eligible students in the population. We summed the weights of those students for whom sex was predicted correctly and divided this by the sum of all students weights for a given country.

This approach can be illustrated with the 2015 data from the U.S. Given the aim to determine the sex of students in the U.S. from binomial logistic regression models based on data from other countries, we first calculated the regression coefficients for all countries separately. In these regression models, the dependent variable is sex and the independent variables are the scores on the reading, science, mathematics, and the six science attitude variables used in the 2015 PISA, such as "joy in science". For example, we calculated the binomial regression model for Germany using German data, the model for the U.K. using U.K. data, and so on for all countries except the U.S. (for 2015, $N = 55 - 1$, Table 1). Because of the involvement of plausible variables, we calculated a separate logistic regression for each plausible variable set. Next, we used the calculated regression models of all 54 countries (excluding the U.S.) to predict student sex in the U.S. (again, separately for each plausible variable set). In other words, we applied, for example, the German set of regression coefficients to the U.S. student data to predict the sex of U.S. students. Doing this, one finds that the model of Israel was the poorest in predicting the sex of U.S. students (62% correctly classified) and the model of Luxembourg the best (65% correctly classified). On average, the sex of 63% of U.S. students was classified correctly. We call this the universal prediction of student sex for the U.S.

In comparison, using the U.S. regression model based on U.S. student data, the sex of 65% of U.S. students was classified correctly. We call this the "local" prediction of student sex. Hence, the difference between the local (U.S.) model's prediction and the average universal prediction was two percentage points. This difference reflects country-specific effects and measurement error. The country-specific effects should provide an estimate of unique cultural or educational influences on sex differences in academic achievement and attitudes.

The extent of correct classification can also be expressed as Cohen's *d*, by calculating the Cohen's *d* for the log odds of being male (reference value male is one, female is zero) for each of the students. For the U.S., Cohen's *d* of the log odds of sex based on the U.S.'s own regression model (i.e., local) is d=0.77. Note again that this is identical to the multivariate Mahalanobis distance (a.k.a. multivariate *d*). Similarly, the universal effect for the U.S. (63% correct classification) corresponds to d=0.69.

Further, we also carried out the same calculations for the three individual achievement variables and the

six science attitude variables separately. That is, how well can one predict the sex, for example, of U.S. students using the logistic regression models of other countries using only one achievement or attitude variable (e.g., just "reading", or just "joy in science"). Using this methodology, the variable “broad interest in science” is the most predictive. That is, using the logistic regression models of other countries (each predicting student sex based on only "broad interest in science"), 57% of U.S. students' sex can be predicted correctly (this corresponds to d=0.3). Note that this is lower than the 63% (corresponding to d=0.69) for the universal multivariate method.

The PISA supplies a measure called student weight, which is the number of students that are represented by each assessed student. This measure was not included in our calculations of the logistic regression models, but it was applied in the subsequent classification of each individual. The reason is that student weights varied much more for some countries than for others. For example, for the U.S. 2015 PISA, 5,712 students were sampled from a population of 3.5 million 15 year olds. Weights in the U.S. data set vary between 29.9 and 2,160.9. In other words, some sampled students represented a relatively large portion of other students within the U.S., which makes it less likely that such cases are fully representative. In other countries, the weights are far less variable, simply because more students were sampled and because the overall population is smaller. For example, in Luxembourg, weights varied between 1.0 and 1.3. Because of this, we treated all students the same in the creation of our logistic regression models (i.e., all students were given weight 1.0). Note that the aim of the classifier creation (i.e., logistic regression in our case) is agnostic of the algorithm underlying the classifier – in that sense, we just worked without weights because that resulted in a better classification. However, when applying the models to classify students, we used the PISA weights because in this way our prediction of success for any particular country matches the sample (including the fact that some students are more representative than others for the specific sample).

Results

Across countries, the mean local prediction (correct classification) of student sex is 73.2%, 73.6%, and 66.6% for the three PISA cycles 2009, 2012, and 2015, respectively (i.e., the percentage of students whose sex could be predicted based on their achievement and attitude variables from their own country). The mean universal prediction of student sex for the three cycles is 71.3%, 70.7%, 64.8%, respectively. Thus, the average universal classification success is consistently slightly lower than the classification based on the local model (Figure 1). In other words, one can classify the sex of students in country X slightly better with the regression model of country X (i.e., local) than using the regression models of all other countries (i.e., universal), as confirmed with a paired t-test on the local model’s and universal classification models (p<.001 for each cycle). The lower classification of the universal prediction is expected, but is only a few percentage points lower than the local models: the associated standardized regression coefficients of all models for each PISA cycle are in the supplementary online material.

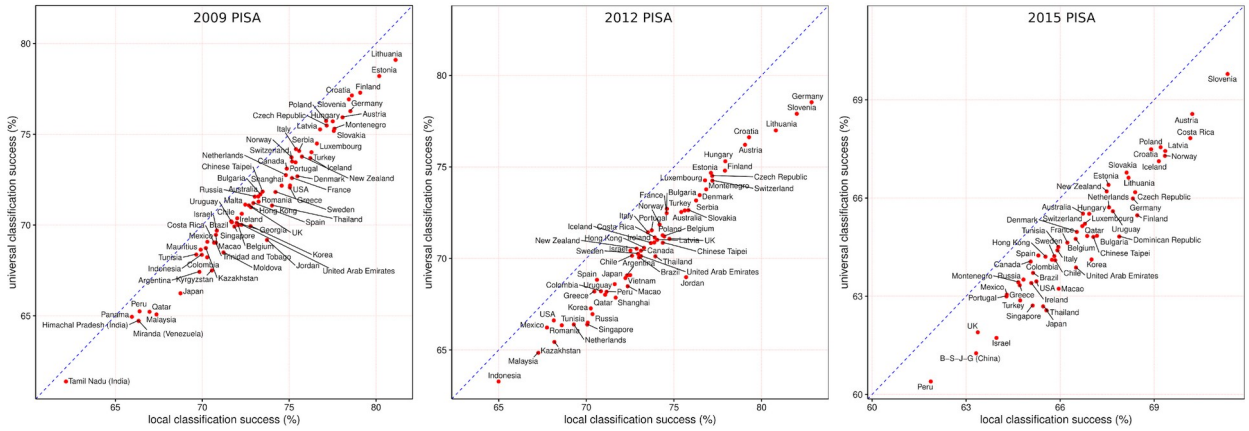

Figure 1: Success of classifying student sex based on achievement and attitude patterns for the three PISA cycles. The dashed blue line indicates identity (x=y). Note that local classification is always slightly more successful than universal classification.

That the sex of students can be predicted based on the regression models from other countries is consistent with a universal pattern of sex differences in academic strengths and weaknesses. As described in the Methods, Cohen's *d* of the log odds of a specific sex provides an estimate of the magnitude of these sex differences. The average effect sizes of the local models in 2009, 2012, and 2015, were 1.24, 1.26, and 0.84, respectively. The average effect sizes of the universal models in 2009, 2012, and 2015 were 1.13, 1.10, and 0.75.

Next, we aimed to distinguish between the contributions of the achievement variables and the contributions of the attitude variables to classification success. To do so, we repeated the same analyses as above, but first only including the three achievement variables and second only including the attitude variables. Using only the three achievement variables, the average successful universal classification varies from 64% to 70%, as compared to 58% to 62% when using only the attitude variables (for data and plots, see supplementary online material). In each of the three PISA cycles, the classification success based on attitudes-only data was lower than the classification success based on achievement-only data (paired *t*-tests, $p<.001$). Finally, the classification success based on all variables (achievement and attitudes) was better than the success based on achievement only (paired *t*-tests, $p<.001$).

The sex of students in some countries can be predicted better than the sex of students in other countries. The variability in the success of universal classification in achievement-only models correlates with the universal classification success in attitudes-only models. That is, when student sex for a country can be well predicted based on achievement-only data, sex can be predicted based on attitudes-only data ($r=.65$, $r=.66$, $r=.32$ in 2009, 2012, and 2015, respectively). It should be noted that the variability in these three correlations from the three different years is partially due to the fact that quite different attitude variables were used. In other words, some attitude variables will correlate better with achievement than others (see Appendix A for a detailed list of attitude variables used).

Next, to demonstrate the utility of our approach we compared classification success using the multivariate approach with the success of individual variables. To do so, we calculated the universal prediction for each individual variable (Table 2).

Year	Variable	Success (%)	*D*	International sex differences (averaged *d*)	International sex differences (range *d*)
2009					
	Mathematics	51.67	0.09	0.10 (0.11)	-0.14 — 0.44
	Reading	58.43	0.43	-0.44 (0.13)	-0.72 — -0.11
	Science	50.44	0.00	-0.04 (0.13)	-0.4 — 0.27
	Enjoyment of reading	62.49	0.61	-0.61 (0.17)	-1.05 — -0.16
	Library use	53.72	0.22	-0.22 (0.11)	-0.55 — -0.01
	Online reading	51.15	0.04	0.06 (0.1)	-0.18 — 0.28
	Diversity of reading	53.41	0.19	-0.21 (0.13)	-0.49 — 0.08
2012					
	Mathematics	52.66	0.12	0.09 (0.12)	-0.27 — 0.35
	Reading	58.14	0.42	-0.44 (0.14)	-0.9 — -0.22
	Science	50.79	0.01	-0.03 (0.13)	-0.53 — 0.23
	Mathematics Self Concept	56.3	0.32	0.31 (0.14)	-0.04 — 0.66
	Interest in mathematics	54.37	0.21	0.19 (0.12)	-0.12 — 0.51
	Instrumental motivation for mathematics	53.04	0.14	0.15 (0.14)	-0.17 — 0.56
	Mathematics behavior	57.85	0.36	0.29 (0.1)	0.04 — 0.51
	Mathematics anxiety	52.81	0.14	-0.23 (0.16)	-0.51 — 0.24
	Mathematics self-efficacy	56.2	0.33	0.28 (0.12)	0 — 0.53
	Mathematics intentions	55.25	0.27	0.28 (0.17)	-0.12 — 0.66
	Mathematics work ethic	50.84	0.01	-0.09 (0.09)	-0.27 — 0.12
	Attributions to failure in mathematics	51.23	0.04	-0.10 (0.12)	-0.31 — 0.21
	Subjective norms in mathematics	52.46	0.13	0.11 (0.09)	-0.13 — 0.33
2015					
	Mathematics	51.93	0.09	0.05 (0.11)	-0.18 — 0.29
	Reading	55.25	0.26	-0.32 (0.13)	-0.82 — -0.09
	Science	51.18	0.04	-0.01 (0.13)	-0.48 — 0.26
	Interest in broad science topics	56.09	0.23	0.22 (0.1)	-0.02 — 0.45
	Science activities	56.79	0.33	0.33 (0.08)	0.15 — 0.59
	Joy of science	51.26	0.02	0.05 (0.15)	-0.3 — 0.47
	Science self-efficacy	51.55	0.09	0.10 (0.11)	-0.16 — 0.34
	Instrumental motivation for science	50.19	0.00	0.00 (0.1)	-0.19 — 0.27
	Epistemological beliefs	51.07	0.02	-0.06 (0.09)	-0.4 — 0.12

Table 2: Success rate of classifying sex based on individual variables and sex differences academic achievement and attitudes. For each variable, the success of the classification based on

that variable is reported, as well as the associated multivariate *D*. We also report the international average of the actual sex difference (*d*) with standard deviation in parenthesis, as well as the range in *d* of across countries. Negative values indicate that boys have lower scores than girls.

In summary, for the 2009 data set, the reading comprehension measure alone was the best universal predictor among the achievement variables and correctly classified the sex of 58.4% of students (*d*=0.43). The variable "joy in reading" was the best predictor among the attitude variables and correctly classified the sex of 62.5% of students (*d*=0.61). For 2012, the variable "reading comprehension" was again the best predictor among the achievement variables (58%, *d*=0.42), whereas the variable "math behavior" (see Methods for description) was the best predictor among the attitude variables (57.9%, *d*=0.36). For 2015, the variable "reading comprehension" was once again the best predictor among the achievement variables (55.2%, *d*=0.26), and the variable "science activities" among the attitude variables (56.8%, *d*=0.33). Remember that the multivariate universal success rates, as reported before and repeated here, were considerably higher for each cycle (73.2%, 73.6%, and 66.6%, with effect sizes in the log odds *d*=1.13, *d*=1.10, and *d*=0.75, for the 2009, 2012, and 2015 PISA cycles, respectively).

Finally, we assessed the extent to which universal classification success relates to measures of economic equality (Gini) and empowerment (GGGI). As predicted, higher classification success was associated with higher levels of women's economic and political empowerment and lower levels of income inequality (Table 3). In other words, the global pattern of sex differences is stronger in more egalitarian countries.

PISA Cycle	GGGI	Gini
2009	$r(60) = .37$, $p = .003$	$r(54) = -.56$, $p < .001$
2012	$r(54) = .29$, $p = .029$	$r(51) = -.45$, $p = .001$
2015	$r(49) = .36$, $p = .011$	$r(51) = -.45$, $p = .001$

Table 3: Pearson correlations between the universal prediction of student sex and equality indices for each PISA cycle. Higher scores on the GGGI indicate higher participation of women in politics and the economy, and higher scores on the Gini indicate a more unequal distribution of household income.

Discussion

We show for the first time that student sex can be reliably predicted by a combination of achievement and attitude variables in all assessed countries and regions in three large, international data sets. Critically, 69% of students can be correctly classified as boys or girls based on academic patterns derived from other countries, which is analogous to an average effect size (*d*) of one standard deviation. Moreover, student sex can be predicted much better by the pattern of academic strengths and weaknesses and attitudes than by any individual variable. In combination, the results indicate that there is a pattern of academic competencies and attitudes that is sex-specific and universal (i.e., consistent across countries).

Further, the extent to which student sex can be correctly classified based on models from other countries correlates with economic and social equality: as the sociopolitical and economic equality of a country

increases, their students are more likely to show universal sex-typical patterns of academic achievement and attitudes. Similar results have been found for the Big Five personality traits, which are more strongly expressed in more egalitarian countries (Costa et al., 2001; Schmitt et al., 2008). Likewise, Stoet and Geary (2018) showed that sex differences in engagement with science, technology, engineering, and mathematics (STEM) are more strongly expressed in countries with higher levels of women's participation in the economy and politics.

Theoretical implications

Our work reveals three major findings, each with implications for theories about sex differences in cognitive abilities and attitudes. First, sex differences in the pattern of academic achievement and attitudes are larger than suggested by the assessment of single domains, such as mathematics achievement. For example, the mean sex difference in mathematics achievement (averaged across countries) has been estimated to be around one tenth of a standard deviation, albeit with considerable variation between countries (Else-Quest et al., 2010; Hyde et al., 1990; Hyde & Mertz, 2009). The present study and earlier ones (Stoet & Geary, 2013, 2015, 2018) indicate that this particular (or any other) single-variable sex difference might not be meaningful outside of the context of the overall pattern of academic abilities and attitudes. This is, in part, because students' decisions about college and career paths are based on relative academic strengths in combination with their interests (Lauermann et al., 2017; Stoet & Geary, 2018), which renders theories about sex differences in pursuit of one path or another based on a single domain (e.g., mathematics achievement) incomplete.

This finding is consistent with other studies that reveal larger sex differences when a pattern of related constructs, such as different dimensions of personality, is considered rather than only a single dimension (Conroy-Beam & Buss, 2017; Del Giudice, 2009; Del Giudice et al., 2012). Our research confirms their conclusion that many sex differences in multivariate data sets are large, and extends it in two novel ways. First, one of the criticisms of Del Giudice and colleagues' (2009, 2012) approach was the reliance on self-assessed traits (Hyde, 2012). The argument is that self-assessed traits are more sex-typed because responses on each of the assessed dimensions is influenced by gender stereotypes. Our current study reveals the same large sex differences for achievement tests. In fact, sex was more accurately predicted by achievement than by self-reported attitudes even though the latter should be particularly prone to stereotyped beliefs (and hence particularly large according to the criticism).

Our second major finding, perhaps even more important than the first, is the identification of a universal pattern, whereby the academic and attitude patterns that predict student sex in Estonia, for instance, accurately predict student sex in other regions of the world, including countries in North America, South America, Asia, North Africa, the Middle East, and Oceania. The importance of this universal effect is highlighted by the finding that the models for individual countries were not much better at predicting the sex of their own students than were the models derived from other countries. Given the data used, we cannot determine the reason for this universal pattern with certainty, but the results narrow the range of possibilities. Either the social conditions that cause sex differences (e.g., sex-typed academic stereotypes) are the same throughout the world, or there are biologically influenced sex differences in the competencies (e.g., language and spatial abilities) and interests that support academic achievement and associated achievement attitudes (Geary, 1996, 2007; Su et al., 2009). Any such biologically influenced sex differences could, of course, result in universal stereotypes, but in this case the stereotypes reflect the observation of differences and not the creation of them (Jussim et al., 2016). Given the considerable international variation in social conditions, it seems rather unlikely that all countries will produce the same socially-derived sex differences. On the other hand, the possibility of more inherent sex differences influencing the expressions of academic skills and interests seems most likely, given the shared biology

(Geary, 1996).

One counterargument against the conclusion that biology drives the universal effect is that some very basic physical sex differences (e.g., the larger size of men and pregnancy in women) channel boys and girls into different economic and cultural niches (Eagly, 1987; Wood & Eagly, 2002); specifically, an agentic orientation associated with a focus on economic success for men and a communal orientation focused on the care of children for women. On the basis of this type of division of labor, societies develop socially-enforced norms regarding the behavior of individuals who occupy these niches and these norms in turn result in sex differences in a variety of psychological domains. However, such a process is inconsistent with our finding that the pattern of sex differences in academic achievement and attitudes is more sex-typed in developed countries with diverse economic niches and alternatives to traditional maternal care of children. In fact, cross-cultural studies of child rearing indicate that in comparison to less developed countries with a clear division of labor, parents in developed countries encourage the academic achievement of girls and some level of economic independence (Low, 1989). More fundamentally, the agentic and communal social behaviors ascribed to boys and men and girls and women, respectively, applies to all mammals (Geary, in press). We would like to add, however, that it is impossible to fully separate contributions of biological and social factors (Miller & Halpern, 2014). Our point is not that social factors do not play a role – they clearly do; our point is that biology also plays a role, which is often omitted from influential studies discussing international variation in sex differences (e.g., Else-Quest et al., 2010; Hyde & Mertz, 2009).

Finally, our third major finding that the universal predictability is larger in more egalitarian countries corroborates earlier findings of more clearly expressed sex differences in these countries (Costa et al., 2001; Falk & Hermle, 2018; Mac Giolla & Kajonius, 2018; Schmitt et al., 2008, 2017; Stoet & Geary, 2018). This finding is inconsistent with the idea that a reduction of gender stratification (e.g., segregated work environments) will lead to a reduction of sex differences in psychological traits (Eagly & Wood, 1999) and in academic outcomes (e.g., Else-Quest et al., 2010; Spelke, 2005).

At this point, it is impossible to determine with certainty the cause of the correlation between scores of economic and political equality and sex differences in psychological traits. One possible explanation is that in more egalitarian countries, education is less an instrument to overcome poverty and to improve the quality of life (Stoet & Geary, 2018). That is, the risk of economic difficulties associated with choosing a career without good prospects of a reliable income will funnel students of both sexes into certain types of employment (e.g., a degree in computer science instead of a degree in medieval literature). With the lessening of these risks, we expect that students' interests and academic strengths will more strongly influence their educational and occupational choices that in turn more strongly reveal any underlying sex differences. In order to fully test such a model, future research should focus on a more direct relation between educational achievement and the socioeconomic aspirations of parents and students. For example, it might be the case that sex differences in mathematics achievement are reduced in countries where parents or students believe that mathematics is absolutely necessary for a well-paid career.

A major question is what (biological) mechanism causes the universal and inevitable existence of sex differences in the pattern educational traits. As we described in the introduction and as an example, there are well-documented sex differences in spatial and language abilities (among others) that are correlated with mathematics and reading achievement, respectively (Bus & Van Ijzendoorn, 1999; Geary et al., 2000). These fundamental differences contribute to at least some proportion of the sex difference in some mathematics domains and in ease of learning to read. In other words, academic learning is built upon universal cognitive abilities (e.g., language) and any sex differences in these universal abilities likely contribute to sex differences in academic domains (Geary, 2007). These academic sex differences in turn

could take on a life of their own, whereby students invest more in the development of academic competencies in areas that are the easiest for them. Over time, any such investments could exaggerate sex differences in these academic areas and likely influence related attitudes, such as enjoyment of reading for pleasure. More likely than not, there are other sex differences, including interest in people versus things, that influence academic and occupational development (Su et al., 2009). The details remain to be worked out, but this type of process would result in universal sex differences in academic abilities and attitudes and a stronger expression of these in societies in which student have more control (e.g., elective courses in middle school and high school) over their academic development.

References

Archer, J. (2019). The reality and evolutionary significance of human psychological sex differences. *Biological Reviews*, *94*(4), 1381–1415. https://doi.org/10.1111/brv.12507

Asperholm, M., Nagar, S., Dekhtyar, S., & Herlitz, A. (2019). The magnitude of sex differences in verbal episodic memory increases with social progress: Data from 54 countries across 40 years. *PLoS ONE*, *14*(4). Scopus. https://doi.org/10.1371/journal.pone.0214945

Bramble, M. S., Lipson, A., Vashist, N., & Vilain, E. (2017). Effects of chromosomal sex and hormonal influences on shaping sex differences in brain and behavior: Lessons from cases of disorders of sex development. *Journal of Neuroscience Research*, *95*(1–2), 65–74. https://doi.org/10.1002/jnr.23832

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., & Linney, A. (1993). Sex discrimination: How do we tell the difference between male and female faces? *Perception*, *22*(2), 131–152. Scopus. https://doi.org/10.1068/p220131

Bus, A. G., & Van Ijzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology*, *91*(3), 403–414. Scopus. https://doi.org/10.1037/0022-0663.91.3.403

Chekroud, A. M., Ward, E. J., Rosenberg, M. D., & Holmes, A. J. (2016). Patterns in the human brain mosaic discriminate males from females. *Proceedings of the National Academy of Sciences*, *113*(14), E1968–E1968. https://doi.org/10.1073/pnas.1523888113

Conroy-Beam, D., & Buss, D. M. (2017). Euclidean distances discriminatively predict short-term and long-term attraction to potential mates. *Evolution and Human Behavior*, *38*(4), 442–450. https://doi.org/10.1016/j.evolhumbehav.2017.04.004

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

Cotton, S., Fowler, K., & Pomiankowski, A. (2004). Do sexual ornaments demonstrate heightened condition-dependent expression as predicted by the handicap hypothesis? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1541), 771–783. https://doi.org/10.1098/rspb.2004.2688

Dean, D. C., Planalp, E. M., Wooten, W., Schmidt, C. K., Kecskemeti, S. R., Frye, C., Schmidt, N. L., Goldsmith, H. H., Alexander, A. L., & Davidson, R. J. (2018). Investigation of brain structure in the 1-month infant. *Brain Structure and Function*, *223*(4), 1953–1970. https://doi.org/10.1007/s00429-017-1600-2

Del Giudice, M. (2009). On the Real Magnitude of Psychological Sex Differences. *Evolutionary Psychology*, *7*(2), 147470490900700220. https://doi.org/10.1177/147470490900700209

Del Giudice, M. (2013). Multivariate Misgivings: Is D a Valid Measure of Group and Sex Differences? *Evolutionary Psychology*, *11*(5), 147470491301100500. https://doi.org/10.1177/147470491301100511

Del Giudice, M., Booth, T., & Irwing, P. (2012). The Distance Between Mars and Venus: Measuring

Global Sex Differences in Personality. *PLOS ONE*, *7*(1), e29265. https://doi.org/10.1371/journal.pone.0029265

Del Giudice, M., Lippa, R. A., Puts, D. A., Bailey, D. H., Bailey, J. M., & Schmitt, D. P. (2016). Joel et al.'s method systematically fails to detect large, consistent sex differences. *Proceedings of the National Academy of Sciences*, *113*(14), E1965–E1965. https://doi.org/10.1073/pnas.1525534113

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. https://www.scholars.northwestern.edu/en/publications/sex-differences-in-social-behavior-a-social-role-interpretation

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, *54*(6), 408–423. https://doi.org/10.1037/0003-066X.54.6.408

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. https://doi.org/10.1037/a0018053

Escorial, S., Román, F. J., Martínez, K., Burgaleta, M., Karama, S., & Colom, R. (2015). Sex differences in neocortical structure and cognitive performance: A surface-based morphometry study. *NeuroImage*, *104*, 355–365. https://doi.org/10.1016/j.neuroimage.2014.09.035

Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412), eaas9899. https://doi.org/10.1126/science.aas9899

Geary, D. C. (in press). *Male, Female: The Evolution of Human Sex Differences* (Third edition). American Psychological Assoc.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, *19*(2), 229–247. https://doi.org/10.1017/S0140525X00042400

Geary, D. C. (2007). Educating the evolved mind: Conceptual foundations for an evolutionary educational psychology. In J. S. Carlson & J. R. Levin (Eds.), *Educating the evolved mind* (pp. 177–202). Information Age.

Geary, D. C. (2015). *Evolution of vulnerability: Implications for sex differences in health and development*. Scopus. https://doi.org/10.1016/C2014-0-00387-5

Geary, D. C. (2016). Evolution of Sex Differences in Trait- and Age-Specific Vulnerabilities. *Perspectives on Psychological Science*, *11*(6), 855–876. https://doi.org/10.1177/1745691616650677

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex Differences in Spatial Cognition, Computational Fluency, and Arithmetical Reasoning. *Journal of Experimental Child Psychology*, *77*(4), 337–353. https://doi.org/10.1006/jecp.2000.2594

Hag, K. (2002). Gender and Mathematics Education in Norway. In G. Hanna (Ed.), *Towards Gender Equity in Mathematics Education: An ICMI Study* (pp. 125–137). Springer Netherlands. https://doi.org/10.1007/0-306-47205-8_9

Halpern, D. F. (2011). *Sex Differences in Cognitive Abilities: 4th Edition* (4 edition). Psychology Press.

Hyde, J. S. (2005). The gender similarities hypothesis. *The American Psychologist*, *60*(6), 581–592. https://doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S. (2012). *The Distance Between North Dakota and South Dakota. Reader Comments on Del Giudice, Booth, & Irwing (2012)*. https://journals.plos.org/plosone/article/comment?id=info%3Adoi/10.1371/annotation/2aa4d091-db7a-4789-95ae-b47be9480338

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139–155. https://doi.org/10.1037/0033-2909.107.2.139

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*(22), 8801–8807. https://doi.org/10.1073/pnas.0901265106

Jahanshad, N., & Thompson, P. M. (2017). Multimodal neuroimaging of male and female brain structure in health and disease across the life span. *Journal of Neuroscience Research*, *95*(1–2), 371–379. https://doi.org/10.1002/jnr.23919

Johnson, E. S. (1984). Sex differences in problem solving. *Journal of Educational Psychology*, *76*(6), 1359–1371. https://doi.org/10.1037/0022-0663.76.6.1359

Jussim, L., Crawford, J. T., Anglin, S. M., Chambers, J. R., Stevens, S. T., & Cohen, F. (2016). Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. In *Handbook of prejudice, stereotyping, and discrimination, 2nd ed* (pp. 31–63). Psychology Press.

Lauermann, F., Tsai, Y.-M., & Eccles, J. S. (2017). Math-related career aspirations and choices within Eccles et al.'s expectancy-value theory of achievement-related behaviors. *Developmental Psychology*, *53*(8), 1540–1559. https://doi.org/10.1037/dev0000367

Lewis, A. B. (1989). Training Students to Represent Arithmetic Word Problems. *Journal of Educational Psychology*, *81*(4), 521–531. Scopus. https://doi.org/10.1037/0022-0663.81.4.521

Lippa, R. A. (2005). *Gender, nature, and nurture, 2nd ed*. Lawrence Erlbaum Associates Publishers.

Low, B. S. (1989). Cross-cultural patterns in the training of children: An evolutionary perspective. *Journal of Comparative Psychology (Washington, D.C. : 1983)*, *103*(4), 311–319. Scopus. https://doi.org/10.1037/0735-7036.103.4.311

Mac Giolla, E., & Kajonius, P. J. (2018). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology: Journal International De Psychologie*. https://doi.org/10.1002/ijop.12529

Majeres, R. L. (2007). Sex differences in phonological coding: Alphabet transformation speed. *Intelligence*, *35*(4), 335–346. Scopus. https://doi.org/10.1016/j.intell.2006.08.005

Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences*, *18*(1), 37–45. https://doi.org/10.1016/j.tics.2013.10.011

OECD. (2009). *PISA Data Analysis Manual*. OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. OECD Publishing.

OECD. (2014). *PISA 2012 Technical Report*. OECD Publishing.

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.

OECD. (2018). *PISA Test Questions*. https://www.oecd.org/pisa/pisaproducts/pisa-test-questions.htm

Ontario Ministry of Education. (2004). *Me read? No way! A practical guide to improving boys' literacy skills*. http://www.edu.gov.on.ca/eng/document/brochure/meread/meread.pdf

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, *74*(4), 445–458.

Rosenblatt, J. D. (2016). Multivariate revisit to "sex beyond the genitalia." *Proceedings of the National Academy of Sciences*, *113*(14), E1966–E1967. https://doi.org/10.1073/pnas.1523961113

Schmitt, D. P., Long, A. E., McPhearson, A., O'Brien, K., Remmert, B., & Shah, S. H. (2017). Personality and gender differences in global perspective. *International Journal of Psychology*, *52*(S1), 45–56. https://doi.org/10.1002/ijop.12265

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*(1), 168–182. https://doi.org/10.1037/0022-3514.94.1.168

Spelke, E. S. (2005). Sex Differences in Intrinsic Aptitude for Mathematics and Science?: A Critical Review. *American Psychologist*, *60*(9), 950–958. https://doi.org/10.1037/0003-066X.60.9.950

Stoet, G., & Geary, D. C. (2013). Sex Differences in Mathematics and Reading Achievement Are Inversely Related: Within- and Across-Nation Assessment of 10 Years of PISA Data. *PLOS ONE*, *8*(3), e57988. https://doi.org/10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence*, *48*, 137–151.

https://doi.org/10.1016/j.intell.2014.11.006
Stoet, G., & Geary, D. C. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science*, *29*(4), 581–593. https://doi.org/10.1177/0956797617741719
Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*(6), 859.
van der Linden, D., Dunkel, C. S., & Madison, G. (2017). Sex differences in brain size and general intelligence (g). *Intelligence*, *63*, 78–88. https://doi.org/10.1016/j.intell.2017.04.007
Wang, M.-T., & Degol, J. L. (2017). Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educational Psychology Review*, *29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x
Wood, W., & Eagly, A. H. (2002). A cross-cultural analysis of the behavior of women and men: Implications for the origins of sex differences. *Psychological Bulletin*, *128*(5), 699–727. https://doi.org/10.1037/0033-2909.128.5.699
World Economic Forum. (2006). *Global Gender Gap Report 2006*. World Economic Forum.
World Economic Forum. (2012). *Global Gender Gap Report 2012*. World Economic Forum.

Acknowledgments

We would like to thank the anonymous reviewers and editor for the helpful comments and suggestions on a draft of this manuscript.

Appendix A

Detailed description of the attitude variables. Note that each PISA cycle focuses on one of the three domains (mathematics, reading, and science). In practice, this means that most of the attitude variables are related to the focused domain. For example, in the 2009 PISA cycle, the attitude questions were about reading motivation and behavior, whereas the 2012 PISA cycle's questions were about mathematics motivation and related behavior.

From the 2009 PISA, which focused on reading attitudes, we included the following constructs, as provided by PISA (for details see OECD, 2012). Note that for each construct, PISA provides for each student a standardized score based on their own item response model.

1. Diversity in reading material, based on the question "How often do you read these materials because you want to?" with 5 different types of reading material to score, namely "Magazines", "Comic books", "Fiction (novels, narratives, stories"), "Non-fiction books", and "Newspapers". There were five response categories ("never or almost never", "a few times a year", "about once a month", "several times a month", "several times a week").
2. Enjoyment of reading based on 11 items, namely "I read only if I have to", "Reading is one of my favourite hobbies", "I like talking about books with other people", "I find it hard to finish books","I feel happy if I receive a book as a present", "For me, reading is a waste of time", "I enjoy going to a bookstore or a library", "I read only to get information that I need", "I cannot sit still and read for more than a few minutes","I like to express my opinions about books I have read","I like to exchange books with my friends". Each item was responded to with one of four response categories ("Strongly disagree", "disagree", "agree", "strongly agree").
3. Library use, based on a scoring of frequency for 7 different activities in libraries, namely

“Borrow books to read for pleasure”, “Borrow books for school work”, “Work on homework, course assignments or research papers”, “Read magazines or newspapers”, “Read books for fun”,”Learn about things that are not course-related, such as sports, hobbies, people or music”, “Use the Internet”. Each item was responded to with one of 5 response categories (“never”, “a few times a year”, “about once a month”, “several times a month”, “several times a week”).

4. Online reading. The question "How often are you involved in the following reading activities?" was answered for 7 different activities, namely “Reading emails”,“Chat online (e.g., MSN®)”,“Reading online news”,“Using an online dictionary or encyclopaedia (e.g. Wikipedia®)”,“Searching online information to learn about a particular topic”,“Taking part in online group discussions or forums”, and “Searching for practical information online (e.g. schedules, events,tips, recipes)”. Each item was responded to with one of 5 response categories (“I don’t know what it is”, “never or almost never”, “several times a month”, “several times a week”, “several times a day”).

From the 2012 PISA, which focused on mathematics attitudes, we included the following constructs (for details see OECD, 2014).

1. Mathematics self concept, based on 5 items in response to the question "Thinking about studying mathematics: to what extent do you agree with the following statements?", namely "I am just not good at mathematics", "I get good grades in mathematics", "I learn mathematics quickly", "I have always believed that mathematics is one of my best subjects", and "In my mathematics class, I understand even the most difficult work". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").
2. Attributions to failure in mathematics, based on 6 items in response to the question "Suppose that you are a student in the following situation: Each week, your mathematics teacher gives a short quiz. Recently you have done badly on these quizzes. Today you are trying to figure out why. How likely are you to have these thoughts or feelings in this situation?". The items were "I’m not very good at solving mathematics problems", "My teacher did not explain the concepts well this week", "This week I made bad guesses on the quiz", "Sometimes the course material is too hard", "The teacher did not get students interested in the material", and "Sometimes I am just unlucky". Each item was responded to with one of 4 response categories (“Very likely”, “Likely”, “Slightly likely”, “Not at all likely”).
3. Subjective norms in mathematics, based on 6 items in response to the question "Thinking about how people important to you view mathematics: how strongly do you agree with the following statements?". The specific items were as follows. "Most of my friends do well in mathematics", "Most of my friends work hard at mathematics", "My friends enjoy taking mathematics tests", "My parents believe it’s important for me to study mathematics", "My parents believe that mathematics is important for my career", "My parents like mathematics". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").
4. Mathematics work ethic based on items in response to the question "Thinking about the mathematics you do for school: to what extent do you agree with the following statements?". The 9 items were "I finish my homework in time for mathematics class.", "I work hard on my mathematics homework", "I am prepared for my mathematics exams", "I study hard for mathematics quizzes", "I keep studying until I understand mathematics material", "I pay attention in mathematics class", "I listen in mathematics class", "I avoid distractions when I am studying mathematics", and "I keep my mathematics work well organised". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").

5. Mathematics intentions, based on 5 item pairs of which respondents chose one, namely "1. I intend to take additional mathematics courses after school finishes" vs "2. I intend to take additional English courses after school finishes"; "1. I plan on majoring in a subject in college that requires mathematics skills" vs "2. I plan on majoring in a subject in college that required science skills"; "1. I am willing to study harder in my mathematics classes than is required" vs "2. I am willing to study harder in my English classes than is required"; "1. I plan on taking as many mathematics classes as I can during my education" vs "2. I plan on taking as many science classes as I can during my education"; "1. I am planning on pursuing a career that involves a lot of mathematics" vs "2. I am planning on pursuing a career that involves a lot of science".
6. Mathematics behavior, based on 8 items around the question "How often do you do the following things at school and outside of school?", namely "I talk about mathematics problems with my friends", "I help my friends with mathematics" "I do mathematics as an extracurricular activity", "I take part in mathematics competitions", "I do mathematics more than 2 hours a day outside of school", "I play chess", "I program computers", and "I participate in a mathematics club". Each item was responded to with one of the following response categories: “Always or almost always”, “Often”, “Sometimes”, “Never or rarely”.
7. Mathematics self-efficacy, based on confidence scoring for 8 mathematics activities, namely "Using a train timetable to work out how long it would take to get from one place to another", "Calculating how much cheaper a TV would be after a 30% discount", "Calculating how many square metres of tiles you need to cover a floor", "Understanding graphs presented in newspapers", "Solving an equation like 3x+5= 17", "Finding the actual distance between two places on a map with a 1:10 000 scale", "Solving an equation like 2(x+3) = (x + 3) (x - 3)", and "Calculating the petrol consumption rate of a car". Each item was responded to with one of the following response categories: “Very confident”, “Confident”, “Not very confident”, “Not at all confident”.
8. Mathematics anxiety, based on 5 items related to the question "Thinking about studying mathematics: to what extent do you agree with the following statements?", namely "I often worry that it will be difficult for me in mathematics classes", "I get very tense when I have to do mathematics homework", "I get very nervous doing mathematics problems", "I feel helpless when doing a mathematics problem", and "I worry that I will get poor grades in mathematics". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").
9. Interest in mathematics, based on 4 items related to the question "Thinking about your views on mathematics: to what extent do you agree with the following statements?", namely "I enjoy reading about mathematics", "I look forward to my mathematics lessons", "I do mathematics because I enjoy it", and "I am interested in the things I learn in mathematics". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").
10. Instrumental motivation for mathematics, based on 4 items related to the question "Thinking about your views on mathematics: to what extent do you agree with the following statements?", namely Making an effort in mathematics is worth it because it will help me in the work that I want to do later on", "Learning mathematics is worthwhile for me because it will improve my career prospects", "Mathematics is an important subject for me because I need it for what I want to study later on", and "I will learn many things in mathematics that will help me get a job". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").

From the 2015 PISA, which focused on science literacy, we included the following constructs (for details

see OECD, 2017).

1. Interest in broad science topics. Respondents were to indicate interest in 5 different items, namely "Biosphere (e.g. ecosystem services, sustainability)", "Motion and forces (e.g. velocity, friction, magnetic and gravitational forces)", "Energy and its transformation (e.g. conservation, chemical reactions)", "The Universe and its history", and "How science can help us prevent disease". Each item was responded to with one of 5 response categories (“not interested”, “hardly interested“, “interested”, “highly interested”, “I don’t know what this is”).
2. Science activities. Respondents indicated how often they engaged in 9 activities, namely "Watch TV programmes about science", "Borrow or buy books on science topics", "Visit websites about science topics", "Read science magazines or science articles in newspapers", "Attend a science club", "Simulate natural phenomena in computer programs/virtual labs", "Simulate technical processes in computer programs/virtual labs", "Visit websites of ecology organisations", and "Follow news of science, environmental, or ecology organisations via blogs and microblogging (e.g. Twitter)". Each item was responded to with one of 4 response categories (“Very often”, “regularly”, “sometimes”, “never or hardly ever”)
3. Joy in science, based on 5 items, namely "I generally have fun when I am learning science topics.","I like reading about science.", "I am happy working on science topics.", "I enjoy acquiring new knowledge about science.", and "I am interested in learning about science.". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").
4. Science self-efficacy. Respondents indicated how easy it would be to do 8 given items on their own, namely "Recognise the science question that underlies a newspaper report on a health issue.", "Explain why earthquakes occur more frequently in some areas than in others.", "Describe the role of antibiotics in the treatment of disease.", "Identify the science question associated with the disposal of garbage.", "Predict how changes to an environment will affect the survival of certain species.", "Interpret the scientific information provided on the labelling of food items.", "Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars.", and "Identify the better of two explanations for the formation of acid rain."). Each item was responded to with one of 4 response categories (“I could do this easily”, “I could do this with a bit of effort”, “I would struggle to do this on my own”, “I couldn’t do this”).
5. Instrumental motivation for science, based on 4 items related to the question "How often do these things happen in your science lessons?" (students could freely choose one specific science subjects they are being taught), namely "The teacher tells me how I am performing in this subject.", "The teacher gives me feedback on my strengths in this science subject", "The teacher tells me in which areas I can still improve", "The teacher tells me how I can improve my performance", and "The teacher advises me on how to reach my learning goals". Each item was responded to with one of 4 response categories (“never or almost never”, “some lessons”, “many lessons”, “every lesson or almost every lesson”).
6. Epistemological beliefs, based on agreement with 6 items, namely "A good way to know if something is true is to do an experiment", "Ideas in science sometimes change", "Good answers are based on evidence from many different experiments", "It is good to try experiments more than once to make sure of your findings", "Sometimes scientists change their minds about what is true in science", and "The ideas in science books sometimes change". Each item was responded to with one of 4 response categories ("Strongly agree", "Agree","Disagree", "Strongly disagree").