Supplementary Information for

Gender stereotypes can explain the gender-equality paradox

Thomas Breda, Elyès Jouini, Clotilde Napp, Georgia Thebault

Thomas Breda
Email: thomas.breda@ens.fr

This PDF file includes:

Materials and Methods
Supplementary Text

Tables S1 to S13

Other supplementary materials for this manuscript include the following:

Databases and codes allowing the replication of the results have been deposited on a third-party server and can be accessed at:
https://www.openicpsr.org/openicpsr/project/123361/version/V1/view

www.pnas.org/cgi/doi/10.1073/pnas.2008704117

Table of Contents

Materials and Methods

Data

The 2012 PISA survey

The Programme for International Student Assessment (PISA) is an every-three-year international survey of 15-year-old students aimed at determining their knowledge and skills in different domains. Students' abilities are assessed in the three curricular domains: mathematics, reading, and science. Students also answer a background questionnaire, seeking information about the students themselves, their homes, and their school and learning experiences. School principals also complete a questionnaire that covers the school system and the learning environment. The assessment does not just ascertain whether students can reproduce knowledge; it also examines how well students can extrapolate from what they have learned and can apply that knowledge in unfamiliar settings, both in and outside of school.

The PISA target population is made up of all students in any educational institution between the ages of 15 years and 3 months and 16 years and 2 months at the time of the assessment. This specific age has been chosen because it is close to the end of compulsory education in most countries. Efforts have been made to insure the absence of cultural or national biases in the test items and in the evaluation of performance.

We analyze data from the PISA 2012 survey (see *(1)* for a summary of the results obtained from this survey). The student data set contains around 485,000 observations, which roughly represent a population of 28 million 15-year-olds attending seventh grade or above in 64 countries, 34 of which belong to the Organisation for Economic Co-operation and Development (OECD) in 2012.

PISA surveys systematically assess students' performance and knowledge in three core subjects: mathematics, reading and science. However, one of the three core subjects is chosen to be covered in greater depth in each survey. In 2012, mathematics literacy is the major subject area, as it was in 2003. This allows us to get more in-depth information on the students' mathematics skills. On top of taking math tests, students fill out a background questionnaire that provides contextual information about themselves, their homes, and their school and learning experiences. The background questionnaire takes 30 minutes to complete and seeks information about students' engagement with and at school in general and engagement with mathematics in particular. It includes questions on students' motivation to succeed in mathematics, the beliefs they hold about themselves as mathematics learners, and their dispositions and behaviors in math-related fields. Of particular interest to us are questions about students' subjective norms in mathematics as well as questions about their self-concept, their anxiety in mathematics, their instrumental or intrinsic motivation for math, their math behaviors, and their intentions to pursue math-related studies and careers (see details below).

Relevant sample size in PISA

The student questionnaire has a rotation design, which means that all students do not answer the same set of questions. The rotated design is such that there are three different forms of the questionnaire, each containing a common part and a rotated part. The common part (which is administered to all students) contains questions about gender, language at home, migrant background, home possessions, parental occupation and education. The rotated part (which is administered to one-third of students) contains questions about attitudinal and other non-cognitive constructs. The rotation design is such that each question/construct is asked in two out of the three different forms of the questionnaire to allow joint analyses of these constructs. This results in responses from two thirds of students per construct in each country, implying that the variables we build at the country level from PISA using these constructs (see below) are based on a total sample of slightly more than 310,000 students.

PISA countries considered

PISA treats Florida, Connecticut and Massachusetts separately from the rest of the United-States because they have a decentralized (specific) management of the PISA survey. This is also the case for the Perm region of the Russian Federation. We have integrated these states/regions to the country they belong to. PISA also considers separately Chinese provinces and cities that claim their independence from China or have strong cultural specificities: Shanghai, Macao, Taiwan and Hong Kong. We have not grouped these cities/regions in a single "China" sample as such grouping may hide strong cultural differences. This leaves us with 64 countries, 34 of them belonging to the OECD.

The PISA Methodology, Plausible Values and Statistical inference with PISA

Details about the PISA methodology can be found in the PISA Technical reports (see *(2)* for 2012), but the following lines give the general idea. PISA adopts the Item Response Theory models and does not provide for each student actual scores in math, reading and science but plausible values. These five plausible values are random numbers drawn from the distribution of scores that could be reasonably assigned to each individual, given his or her answers - that is, the marginal posterior distribution. Any estimation procedure in PISA (for instance mean score of boys) involving students' measured ability in math, reading or science requires the calculation of the targeted statistic for each plausible value (appropriately weighting with the reported student weights) and the final estimate is the arithmetic average of the five estimates obtained. Standard errors are calculated with a replication method that takes into account the stratified two-stage sample design for selection of schools and of students within schools.

Sources of uncertainty in PISA are twofold. First, as explained above, there is some uncertainty on the ability measure of each student and PISA provides five plausible values drawn from a posterior distribution of ability. Second, there is standard sampling error at country-level as performance gaps are not established over the universe of students in a given country. To deal with sampling error, PISA provides 80 alternative sets of individual weights and detailed guideline to use those weights. The computation of corrected standard errors relies on bootstrap techniques: one needs to run the regression of interest for each of the five plausible values, weighting it first by the "true" set of individual weights and then by the 80 alternative sets of weights. The correct point estimate is the average of the 5 regressions ran with the "true" set of weights, while the standard error is computed according to a formula that sums both measurement errors described above.

<u>Country-Level Variables Computed from PISA Items</u>

Students' weights are systematically used to obtain representative statistics. When the estimation of a statistic involves the use of math ability (as a control variable, see below), we use the procedure described in the previous section to deal with plausible values.[1] All variables computed from the PISA survey are standardized to have a mean of 0 and a standard deviation of 1 for each country separately. This transformation is used to obtain gender gaps in the variables of interest for each country that are directly expressed as a fraction of the variable standard deviation in the country. As such, gender gaps are expressed in a similar metric and directly comparable across countries.

1. *Main Country-level measure of Implicit Gender-Math Stereotype (GMS*)

Our country-level measure of interiorized Gender-Math Stereotypes (GMS) is based on the PISA 2012 survey, which includes in the students' questionnaire items about subjective norms and perceived control in math. Our main index relies on national differences between girls and boys in the degree to which they agree with the following assertions:

- "Doing well in math is completely up to me" (*perceived control*, item st43q02, hereafter called B1)
- "My parents believe that math is important for my career" (*subjective norms*, item st35q05, hereafter called B2)

Possible answers to these two statements are 1: Strongly agree, 2: Agree, 3: Disagree and 4: Strongly disagree. This implies that smaller values of B1 and B2 correspond to more perceived control and worse subjective norms.

We consider these items as good proxies for the internalization of gender math-related stereotypes, as they do not express intrinsic motivation or preferences for math, but rather what students perceive as their ability to succeed and norms from significant others.

To construct our measure of interiorized gender math-related stereotypes, we proceed in three steps. First (Step 1), we standardize B1, B2 and plausible values for math performance (pvmath) at the country-level so that their weighted mean (using students' weights provided in PISA to make the country samples representative) is equal to zero and their weighted standard deviation is equal to one. Second (Step 2), we take the arithmetic average of the standardized variables B1 and B2 for every student, and standardize this measure at the country level once again so that its weighted mean is equal to zero and its weighted standard deviation is equal to one. Note that results are qualitatively similar if we omit the standardizations in step 1 and step 2. The objective of the standardizations in step 1 is to give an equal weight to the two items B1 and B2 used to compute GMS. The goal of the second standardization is to compute gender gaps that can be expressed as a proportion of the standard deviation of (B1+B2)/2 in each country.

[1] In practice, this is done with the Stata software using the command "repest".

Finally (Step 3), we regress separately in each country this average on a dummy variable equal to one for female students, controlling for students' ability in math. Our measure is simply the estimated effect of being a female student in this regression, which captures the difference between female and male students with respect to interiorized gender-math stereotypes for each country separately. We call this measure GMS (for "gender-math stereotypes"). The regression is estimated by weighted least squares using students' weights and we use the procedure detailed in the previous section to deal with plausible values for math ability.

2. *Alternative Measures of Gender-Math Stereotypes in 2012*

To verify that our results still hold regardless of the specifications and the type of items chosen in the student questionnaire, we also construct alternative measures of the internalization of gender stereotypes by PISA students:

- *Conservative GMS (GMS2)*: this index is similar to GMS, but controls for both students' performances and declared interest in math (intmat) in step 3 of the algorithm described in the previous section.
- *Gender differences in attribution to failure in mathematics (GMS3)*: we regress an index of perceived self-responsibility for failing in mathematics (Failmat) on a dummy variable equal to one for female students controlling for math performance, and recover the coefficients for each country. The index Failmat is constructed from students' responses to a set of questions examining the following situation: "Suppose you are a student in the following situation: each week, your mathematics teacher gives you a short quiz. Recently you have done badly in these quizzes. Today you are trying to figure out why." The 6 possible responses are the following: "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", "Sometimes I am just unlucky".
- *Gender differences in mathematics self-efficacy in calculating the petrol-consumption rate of a car (GMS4):* the PISA survey has several items measuring students' belief that they can successfully do a range of math tasks, like using a train timetable to work out how long it would take to get from one place to another or calculating how much cheaper a TV would be after a 30% discount. We chose an item about students' confidence in calculating the petrol-consumption rate of a car (item st37q08: "How confident do you feel about having to do the following mathematical task: calculating the petrol consumption rate of a car?"), which is more associated with masculinity. We regress this item on a dummy variable equal to one for female students and controlling for math performance, then recover the coefficients for each country.

3. *Gender gap in intentions to study math*

Our main outcome of interest is the country-level gender gap in intentions to pursue math-related studies or careers. In each country, it is computed as the difference between the weighted mean of the index of mathematics intentions (matintfc) for boys and for girls (using students' weights). The index of mathematics intentions is constructed from a series of five questions that ask students if they are willing (i) to study harder in math versus English/reading courses, (ii) to take additional

math versus English/reading courses after school finishes, (iii) to take a math major versus a science major in college, (iv) to take a maximum number of math versus science classes, and (v) to pursue a career that involves math versus science.
We also resort to actual measures of gender occupational differences. Details about these variables are presented in the next section.

4. *Consistent measures of math attitudes from PISA 2003 and PISA 2012*

To study the evolution of gender-math stereotypes over time, we exploit questions on math attitudes that are available in the two PISA surveys that had a special focus on math and measured these attitudes: PISA 2012 and PISA 2003. Unfortunately, the questions used to build our preferred measures of gender-math stereotypes (see above) are only available in PISA 2012. As a consequence, we use instead a set of questions capturing instrumental motivation for math as they are available in both PISA 2012 and PISA 2003:

- Making effort in math is worthwhile: "Making an effort in mathematics is worth it because it will help me in the work that I want to do later on".
- Mathematic will help me get a job: "I will learn many things in mathematics that will help me get a job."
- Math is important for my future study: "Mathematics is an important subject for me because I need it for what I want to study later on."
- Math is important for my career: "Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>"

Gender gaps in these variables have been computed for years 2012 and 2003, controlling for math performances in the same fashion as gender gaps in B1 and B2.

These measures are likely to be close to students' intentions to study math (which we do not observe in 2003) and they are much less well suited than B1 and B2 to capture how students interiorize stereotypes depending on their gender. Nevertheless, the evolution of the gender gaps in these variables are likely to be correlated with the evolution of Gender-Math Stereotypes between 2003 and 2012 and will be used a proxy for the latter.

<u>Country-level Variables Retrieved from other Data Sources</u>

Our main measure of gender-math stereotypes and our measure of intentions to pursue math-related studies or careers come from PISA2012 (see methods). PISA2012 is also used to construct alternative measures of stereotypes (see above) and student-level control variables used in robustness checks (see below). The other country-level variables used in the study have been retrieved from various data sources and are described below.

1. *Measures of Economic Development*

Human Development Index (HDI)

The Human Development Index (HDI) is a composite statistic of life expectancy, education, and per capita income indicators. A country scores a higher HDI when life expectancy, education level and per capita income is higher.
Source:
https://ourworldindata.org/human-development-index
Values have been taken for year 2012 when available and replaced by the closest value available between 2010 and 2014 when necessary.

Gross Domestic Product (GDP)

GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2010 U.S. dollars.
Source:
https://data.worldbank.org/indicator/NY.GDP.PCAP.KD

2. *Measures of Inequality/equality*

GINI Index

GINI index measures the extent to which the distribution of income among individuals or households within an economy deviates from a perfectly equal distribution where income is defined as household disposable income in a particular year. It consists of earnings, self-employment and capital income and public cash transfers; income taxes and social security contributions paid by households are deducted. A Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.
Source:
GINI index World Bank: https://data.worldbank.org/indicator/SI.POV.GINI. Values have been taken for year 2012 when available and otherwise replaced by the closest value available between 2010 and 2014.
GINI index OECD: https://www.oecd.org/social/income-distribution-database.htm Values have been taken for year 2012 for Japan and New Zealand which are absent from the dataset from the World Bank.[2]

Gender Gap Index (GGI)

The Gender Gap Index, from the World Economic Forum, synthesizes the position of women in any given country by taking into account economic opportunities, economic participation, educational attainment, political achievements, and health and well-being. Larger values point to a better position of women in society.
Source:

[2] The addition of these two countries from an alternative data source has no effect on the paper's main conclusions.

GGI World Bank:
https://tcdata360.worldbank.org/indicators/af52ebe9?country=BRA&indicator=27959&viz=line_chart&years=2006,2018

Values have been taken for year 2012 when available and otherwise replaced by the closest value available between 2010 and 2014.

Coefficient of Human Inequality

The Coefficient of Human Inequality is an average of inequalities in health, education, and income. It is calculated by an unweighted arithmetic mean of estimated inequalities in these dimensions. When all inequalities are of a similar magnitude, the coefficient of human inequality and the overall loss in HDI differ negligibly; when inequalities differ in magnitude, the loss in HDI tends to be higher than the coefficient of human inequality.
Source:
Coefficient of Human Inequality, HDR: http://hdr.undp.org/en/data.
Values have been taken for year 2012 when available and otherwise replaced by the closest value available between 2010 and 2014.

3. *Occupational Segregation*

Share of women among STEM graduates

This variable is taken from the work of Stoet and Geary *(3)*, which relies on data from the UNESCO databases for years 2012 to 2015 (stats.uis.unesco.org). As explained in their corrigendum *(4)*, it is the "propensity of women" to graduate with STEM degrees $a/(a + b)$, where a is the percentage of graduating women who graduate with STEM degrees and b is the percentage of graduating men who graduate with STEM degrees.
In their recent commentary on this matter, Richardson et al. *(5)* argue that this measure is misleading, as it is not appropriate to capture individual STEM preferences. They commend the use of the actual share provided by the UNESCO, consistent with a focus on achieved outcomes. Even though we understand these remarks, we chose to keep Stoet and Geary's measure for two reasons. First, keeping Stoet and Geary's figures is an interesting reference point as we try to develop an alternative explanation to the gender-equality paradox. Second, Stoet and Geary's measure is not affected by imbalances in the numbers of men and women in the population of tertiary education graduates which can mechanically affect the share of women among STEM graduates. Countries that have few women relative to men in tertiary education overall are likely to have few women in STEM as well. This general composition effect is not related to gender segregation across fields of study, and we therefore think that it is more appropriate to use measures of occupational segregation that control for it, as does Stoet and Geary's proposed measure.

Female overrepresentation in humanities

This variable is taken from *(6),* which relies on data from the UNESCO's *Statistical Yearbooks* for years 1995, 1997 and 1998. It is a sex-segregation parameter, which contrasts the female-to-male ratio in humanities to that in the average field of study. More precisely, sex segregation in humanities is defined by $\ln\left(\frac{F_{humanities}}{M_{humanities}}\right) - (\frac{1}{\mathrm{J}})\sum_{1}^{J}\ln\left(\frac{\mathrm{F}_j}{\mathrm{M}_j}\right)$,where F_j and M_j are the number of women and men graduates in each field j, and J is the total number of fields. Humanities encompasses the following subfields of study: education, humanities, art, law, social and behavioral sciences, business, mass communications, home economics, service trades. Negative values of the variable indicate female underrepresentation, and positive values indicate female overrepresentation relative to the other fields of study. Values closer to zero are indicative of greater gender integration.

Female students in Science in tertiary education

This variable is taken from *(7),* which relies on data from the UNESCO databases (stats.uis.unesco.org). It is the percentage of women among individuals enrolled in tertiary science education (2000-2008).

4. *Gender-Science IAT (7)*

The Gender-Science Implicit Association Test (IAT) is a measure of implicit biases relying on the Implicit Association Test. This measure relies on the differential ease to categorize items depending on whether they associate male and science (and female and liberal arts) or female and science (and male and liberal arts). Most people are able to categorize the items faster and more accurately in the former condition (male=science) compared with the latter (female=science), which is taken to reflect stronger associations of science with male than female and therefore possibly an implicit gender-science stereotype. Considering the mean of IAT results for all participants in a given country provides a measure of these implicit biases at the country level. Note that the IAT measure relies on self-selected samples of individuals taking the test online. Following previous work on this database *(7),* we only consider countries for which more than 50 observations are available.

5. *Measures of gender stereotypes and cultural norms (not related to math)*

To broaden the scope of our analysis, we resort to several measures of gender stereotypes and cultural norms. These indexes are not specifically related to mathematics but encompass larger topics about gender norms and their influence on individuals' behaviors. We divided these measures into three categories: i) measures of traditional gender roles, which capture "vertical" stereotypes about the place of men and women in society, ii) measures of essentialist gender norms, which capture "horizontal" stereotypes and relate to the manner in which some personality traits or values are attributed to a specific gender, and iii) liberal values, which capture the extent to which a society values individualism and cultural liberalism.

Traditional gender roles

We build two measures of traditional gender roles based on the International Social Survey Programme (ISSP). The 2012 wave of the survey includes several questions about family and changing gender roles. We compute the share of individuals[3] agreeing to each of the following statement:

- *Women should take care of their home and children*: "A job is all right, but what most women really want is a home and kids"
- *Being a housewife is fulfilling*: "Being a housewife is just as fulfilling as working for pay"

We match these measures with 34 countries of our sample, among which 28 are members of the OECD. The two resulting variables are aimed at capturing traditional gender roles that may be related to the male primacy ideology (see *(8)* or to what we broadly define as "vertical gender stereotypes"), since they tend to reflect the idea that men are superior to women.

We complete these indexes with two measures from the World Value Survey. The first one is the share of individual agreeing to the following statement "University is more important for a boy than a girl" (WVS 2010-2014), which captures how people can conceive different aspirations for men and women. The second one is a subindex called "equality" (WVS 2005-2009), which encompasses questions about gender equality in education, politics and on the labor market. We consider the opposite of this subindex, so that larger values correspond to respondents having more traditional views regarding the roles of women and men (*e.g.* thinking that women do not need to be represented in politics or should not work). These measures are available for respectively 31 and 37 countries in our sample.

Essentialist gender norms

We also consider gender norms that are not directly related to the male-primacy ideology but that imply essentialist differences between women and men. To capture such norms, we consider gender differences in gender-stereotyped personality traits or values.

We first focus on gender differences in self-reported personality traits that can capture systematic differences in how women and men feel they are or should behave. We retrieve two of the measures based on the Revised NEO Personality Inventory data from 26 countries *(9).* The first is the gender gap in feminine extraversion, which is based on five facets (warmth, gregariousness, assertiveness, excitement seeking and positive emotions) and reflects more loving, sociable, submissive, cautious, and cheerful personalities. The second is the gender gap in feminine openness, which is based on four facets (aesthetics, feelings, actions and ideas) and reflects a preference for feelings and novelty over intellectual interests. These two measures are available for 19 countries in our sample.

Men are believed to be more agentic, and women are believed to be more communal. Concerning gender differences in gender-stereotyped values, we consider the *gender gap* (women minus men) in the share of individuals agreeing with the statement that family is important (WVS, 2010-2014, which is available for 31 countries in our sample). This gender gap captures differences between

[3] Considering responses from women only or men only does not alter substantially the conclusions obtained in the paper.

women and men about internalized communality and caring for others, which are stereotyped as feminine.

Liberal values

Liberal values capture the extent to which a society promotes individualism and related cultural norms as core values.

The first measure we use, "Individualism", is based on Hofstede's cultural dimensions *(7).* It captures the degree to which a society is individualistic (as opposed to collectivist), *i.e.* the extent to which individuals are integrated into groups, and how loose are social links. It is available for 40 countries in our sample, among which 21 are members of the OECD.

The second measure we take is the index of emancipative values from the World Value Surveys (WVS, 2005-2009). It is a composite index which combines several sub-indexes about autonomy, equality, choice and voice.

Methods

Country-level analyses

Pairwise correlations and linear regression models

Most of our analyses are conducted at the country level using the variables described above either constructed from PISA or retrieved from other sources. We perform simple pairwise correlations or use non-weighted linear regression models of the type:

$$Segregation_c = \alpha GMS_c + \beta D_c + \varepsilon_c \tag{1}$$

where $Segregation_c$ is a direct or indirect measure of female underrepresentation in math-related fields of study in country c, GMS_c is a measure of Gender-Math Stereotypes in country c, and D_c is a measure of development or equality in country c. Variants of (1) are estimated using only one of the two regressors GMS_c and D_c and the evolutions of α and β across specifications are examined.

Normalizations

We perform two normalizations to ease the interpretation of the results. First, we take the opposite of all variables capturing countries' inequality in order to obtain measures that are increasing with the extent of equality in a country. Similarly, we construct measures of gender segregation across fields of study that are systematically increasing with the actual extent of segregation or female underrepresentation in math.

Second, before any regression estimating a variant of model (1), we standardize all variables entering the model on the regression sample. This allows us to compare the magnitude of the coefficients α and β across specifications as they are expressed in a similar metric. More specifically, α measures by how many standard deviations female underrepresentation in math varies when Gender-Math Stereotypes vary by one standard deviation. Similarly, β measures by how many standard deviations female underrepresentation in math varies when Gender-Math Stereotypes vary by one standard deviation.

Statistical inference

The statistical inference for the country-level regressions does not correct for measurement error in measures of Gender-Math Stereotypes or gender gaps in math-related fields of studies taken from PISA. This choice is driven by the fact that our estimates of the marginal effect of gender-math stereotypes are usually very significant by statistical standards. Measurement error on the left hand side of our regression models is likely to inflate the standard errors of the estimates rather than reducing them, while measurement error on the right hand side would result in attenuation bias *(10).* As a result, accounting for measurement error in country-level variables computed from PISA is likely, if anything to increase both the magnitude of the estimates of β and their statistical significance. Another reason for not doing those corrections systematically is to keep the baseline analysis as simple as possible, so that it can be easily replicated and does not involve heavy bootstrap procedures. A final reason is that the student-level counterparts of the macro-level analyses described in the next section largely confirm that measurement error is not likely to affect our results significantly.

<u>Individual-level analyses</u>

We also provide estimates from individual-level regressions on a sample of about 300,000 students in the 64 countries included in the study in Tables S8 to S11. This allows us to control for unobserved individual-level heterogeneity (at the cost of reducing the sample size) and to correct estimates and standard errors for measurement errors in some of the country-level variables constructed with PISA (see details below). The closest micro-level counterpart to our main cross-country regression model is as follows:

$$Math\_Intentions_{ic} = \delta Girl_{ic} + \alpha'(GMS_c * Girl_{ic}) + \beta'(D_c * Girl_{ic}) + \mu X_{ic} + \gamma_c + \varepsilon_{ic} \qquad (2)$$

where $Math\_Intentions_{ic}$ is the index capturing intentions to study math for student *i* in country *c* (normalized to have a mean of 0 and a standard deviation of 1 in each country), $Girl_{ic}$ is a dummy variable equal to 1 if student *i* in country *c* is a girl, X_{ic} is a vector of control variables (including math ability, interest for mathematics, the level of education of the student's parents, measured both in years and kind of diploma obtained, grade repetition, an index of economic, social and cultural status of the household, a measure of home educational resources, and a measure of attitude towards school, namely how much students think that trying hard at school is important), and γ_c a vector of country fixed effects. For a specific country *c*, γ_c capture the mean of math intentions for boys in the country if the values of GMS_c and D_c were equal to 0 (i.e. the average across the sample of countries as these variables are standardized at the country level) and if the country had the average gender gap in math intentions in the sample. δ captures the average gender gap in math intentions across the whole sample. α' and β' are the micro-level counterparts of the parameters of interest α and β obtained from equation (1). They capture how the gender gap in intentions to study math vary with countries' gender-math stereotypes and development or equality. The magnitude of the macro- and micro-level estimates are however not directly comparable because the standardization of variables used at the macro level cannot be easily reproduced at the micro level. Equation (2) is estimated by weighted least squares using students' weights normalized to sum to one in each country. Such "senate" weights ensure that each country has the same weight in the analysis instead of contributing according to its total population. Standard errors are clustered at the country level, as it is the relevant level of analysis.

We also study the relationship between gender-math stereotypes and countries' development or equality using the following model (instead of a pairwise correlation, as we do at the macro level):

$$\left(\frac{B1+B2}{2}\right)_{ic} = \delta Girl_{ic} + \varphi(D_c * Girl_{ic}) + \mu X_{ic} + \gamma_c + \varepsilon_{ic} \tag{3}$$

where $\left(\frac{B1+B2}{2}\right)_{ic}$ is the average of the responses of student *i* to the questions B1 and B2. Recalling that GMS_c is the gender gap in $\left(\frac{B1+B2}{2}\right)_{ic}$ in country c, we see that φ in model (3) captures how gender-math stereotypes vary with countries development or equality. Senate weights are also used and standard errors are again clustered at the country level.

By considering math intentions at the individual level and clustering standard errors at the country level, model (2) naturally deals with statistical uncertainty related to country-level gender gaps in math intentions. Similarly, by considering B1 and B2 at the individual level, model (3) naturally deals with statistical uncertainty related to the country-level measure of GMS in the analyses linking gender-math stereotypes to countries development or equality.

Supplementary Text

This section discusses in more detail the evidence presented in Table 1 and in the supplementary Tables.

Formal mediation analysis for the main results (Table 1)

We show in Table 1 that the relationship between measures of countries' development or (gender) equality and our indirect measure of female underrepresentation in math-related fields of study (the "gender equality paradox") does not hold anymore once it is controlled for a measure of gender math-related Stereotypes. In contrast, the relationship between gender math-related stereotypes and female underrepresentation in math-related fields is strong and holds both when it is controlled for measures of countries' development or equality and when it is not. We deduced from these results that gender math-related stereotypes mediate the gender equality paradox. To backup more formally this claim, we provide here evidence that our measure of gender math stereotypes is a statistically significant mediator of the gender equality paradox, in the sense that it reduces significantly the marginal effect β of the measure of development (or equality) D_c when it is included in model (1). The common approach to study if β is significantly reduced when the mediator is included is to use a Sobel test *(11-12)*. The test however performs poorly on small samples and we therefore rely instead on bootstrap techniques. Bootstrap results based on 1000 replications show that the marginal effect on female underrepresentation in math-related fields of each of the five measures of development or equality considered in Table 1 is significantly reduced when GMS is added as a control variable (normal-based p-value<0.01 for all five variables). This confirms formally that the gender equality paradox is significantly reduced when controlled for gender math-related stereotypes, so that the latter variable mediates the formal relationship from a statistical point of view.

Math performance and students' perceived control or subjective norms in math (Table S1)

In each country, the gender gaps in B1, B2 and their average are obtained from weighted linear regressions of these variables on a dummy equal to one for female students and a control for students' math performance. Table S1 displays the estimated marginal effect of math ability in these models for every of the 64 countries in our sample. Math performance is significantly and negatively associated with B1 for 22 OECD countries, and positively for 3 OECD countries (the remaining being non-significant).

A similar pattern is observed for B2 and the average of B1 and B2 (used to construct GMS) for which only 4 countries have significant positive values. Altogether, this suggests that higher performance in mathematics is usually associated with negative responses to both statements about perceived control and subjective norms.

Gender gaps in B1, B2 and value of GMS by country (Table S2)

Table S2 displays the gender gaps obtained by linear regression models controlling for math performance (coefficients for controls are presented in Table S1). Gender gaps are generally positive, which implies that female students respond less positively than male students to both statements B1 and B2 about perceived control and subjective norms (as answers that are more positive correspond to lower values of the variable).

The gender gap in B1 conditional on math averages to 0.10 SD over the whole sample, with large variation across countries, from 0.22 SD for Switzerland to -0.13 for Kazakhstan. Gender gaps in B1 are larger for high income countries, about three times larger on average among OECD countries than non-OECD countries. It is positive and significantly different from zero in 29 OECD countries, significantly negative in one country (Mexico) and not significantly different from zero in the four remaining countries. The gender gap in B2 follows a similar pattern, reaching an average 0.15 SD over the 64 countries. Finally, the GMS (gender gap in the average of B1 and B2, see the data section for more details), reaches 0.16 SD on average over the whole sample, with an average twice superior for OECD countries than non-OECD countries. It is significantly positive for 30 OECD countries out of 34, and 19 out of 30 non-OECD countries. There are only two significant negative values, Malaysia and Kazakhstan, though the magnitude of these gaps is relatively low (-0.06 SD). These results show that in most countries, conditional on math performance, female students feel they have less control on their ability to succeed in math than male students, and perceive less than male students that their parents consider math as important for their career.

Correlation between GMS and alternative measures of gender math-related stereotypes (Table S3)

Table S3 displays a country-level correlation matrix of GMS with the gender gaps in B1 and B2 (which both enter the computation of GMS) and three alternative measures of stereotypes associating males to science or mathematics. The two first measures are based on alternative questions taken from the PISA survey and capture gender gaps in mathematics self-efficacy in calculating the petrol-consumption rate of a car and gender gaps in attribution to failure in mathematics (see the SI Data section for details). The third measure is the science-gender IAT that captures individuals' implicit associations between science in general (rather than just math) and males or females.

GMS is strongly and positively correlated with all alternative measures of gender stereotypes. Not surprisingly the correlations are the largest with the gender gaps in B1 and B2 which can be seen as the two subparts of GMS. Interestingly, the correlation between the gender gaps in B1 and B2 is 0.58, showing that that items B1 and B2 are not fully redundant. GMS also exhibits large correlations (0.68 and 0.73) with the two alternative measures of stereotypes constructed from PISA, which are based on gender gaps in attribution to failure in math and self-efficacy to compute the petrol consumption of a car. We see finally that all our measures of gender-math stereotypes computed from PISA have medium to large correlations with the gender-science IAT. Among these measures constructed from PISA, GMS is the one that is the most strongly correlated with the gender-science IAT[4].

Main results for OECD and non-OECD countries separately (Table S4)

Tables S4A and S4B reproduce our main results shown in Table 1 for OECD and non-OECD countries separately.

Panel A of Table S4A shows that the correlations of GMS with measures of development and inequality are positive and significant among OECD countries (only at the 10% level for GDP and the GGI), and quite strong for some measures like the coefficient of human equality (r=0.6). Except for income equality (opposite of the GINI index), the results are quantitatively similar among non-OECD countries (Panel A of Table S4B). However, they are not always statistically significant, possibly due to the limited sample size.

Panels B of Tables S4A and S4B display results for the linear regression of the gender gap in intentions to study math on the GMS or measures of development and equality, or both types of variables. The univariate regressions show that the GMS is strongly associated with the gender gap in intention to study math both in OECD and non-OECD countries (0.6 SD for OECD countries, and between 0.6 and 0.7 SD for non-OECD countries, significant at the 99% level of confidence).

Univariate regressions for measures of development and inequality show only weak evidence of a "Gender Equality Paradox" once we look at OECD and non-OECD countries separately. The coefficients for GDP and GGI are not significant for OECD countries, and only weakly significantly different from zero for the GGI among non-OECD countries. Furthermore, results for the "horse race" models show that the coefficients for measures of development and inequality become systematically non-significant when we include our GMS index. Gender math-related stereotypes appear to mediate entirely the relationship between development and gender gap in intention to study math both among OECD and non-OECD countries.

Relationship between stereotypes and actual gender segregation in higher education (Table S5)

Table 1 shows that the relationship between development, inequality and the gender gap in intentions to study math or pursue math career is mediated by gender math-related stereotypes. We would like to check that similar results are obtained when we consider actual gender segregation in higher education. To this aim, Table S5 reproduces Panel B of Table 1 for three outcomes of occupational segregation used in the literature: i) the propensity of women to graduate with STEM degrees used by *(3)*, ii) the percentage of tertiary students in science who are females used by *(7)*, iii) the index of female overrepresentation in humanities used by *(6)*.

[4] Following *(13)*, we excluded Romania from our sample of analysis for the IAT as it is an outlier. Nonetheless, we computed the same correlations with Romania as a robustness check. Reassuringly, adding Romania leads to very similar results.

The first two panels of the table are based on two alternative measures of women's representation in Science in tertiary education. They support our main conclusions. First, gender-math stereotypes are significantly and negatively associated with the proportion of female students in math-related higher education studies (around -0.6 SD for each regression, $p<0.01$). This effect is about the same magnitude as the effects found in univariate models for gender gap in intentions to study math (Table 1). This suggests that the gender gap in intentions to study math is a good proxy for the actual gender segregation across educational tracks a couple of years later. Second, the coefficients for measures of development and inequality become almost systematically non-significant when we introduce our measure of gender stereotypes in the "horse race" models (for 4 out 5 measures in each of the two panels). Once again, gender-math stereotypes seem to mediate the relationship between development and gender differentiation.

Another way of looking at the underrepresentation in STEM is to consider its counterpart, which is overrepresentation in humanities. Perhaps not surprisingly, the last panel of Table S5 (first row) shows that the relationship between GMS and female overrepresentation in humanities is quantitatively smaller than the relationship between GMS and measures of female underrepresentation in science (around 0.45 SD). This relationship remains however highly significant ($p<0.01$) and it is not strongly affected when a measure of countries' development or equality is included as an additional regressor (with the exception of the measure of "human equality" which lowers more substantially the marginal effect of GMS). In contrast, the relationship between measures of countries development or equality and female overrepresentation in humanities is smaller and not always statistically significant. It becomes close to zero and not significant when GMS is included as an additional regressor. These results show that GMS also mediates the relationship between countries' development and overrepresentation in humanities.

<u>Alternative measure of gender math-related stereotypes that controls for interest in math (Table S6)</u>

As gender differences in preferences for math are likely to be in part shaped by cultural norms and stereotypes, we chose not to control for them when building our main measure of stereotypes GMS. However, to show that the "gender equality paradox" is not driven by gender differences in preferences that may also be of non-cultural origin, we should make sure that our measure of stereotypes does not capture such differences. For this reason, we construct a variant of GMS from the gender gaps in B1 and B2 conditional on both students' ability in math and their declared interest for math. Not surprisingly, controlling for preferences reduces substantially the gender gaps in B1 or B2 and generates a more conservative measure of gender-math stereotypes. The key point, however, is that all our conclusions are maintained when using this more conservative measure (GMS2) instead of GMS. In particular, the correlations between GMS2 and measures of development and inequality are of the same magnitude than the ones displayed in Panel A of Table 1. Similarly, the univariate regression of the GMS on intentions to study math show a coefficient of about 0.6 SD, which is equivalent to results found in our main specification. Horse race regressions lead to the same conclusions: the estimated effect of GMS2 varies from 0.4 to 0.6 SD in the multivariate specifications, and is systematically different from zero while the estimated effect of all measures of development become non-significant, suggesting that GMS mediates the relationship between measures of development and inequality and the gender gap in intentions to study math.

Relationship between alternative measures of gender math-related stereotypes and the gender gap in intentions to study math (Table S7)

As a robustness check for our main specification, we reproduce Panel A and B of Table 1 with alternative measures of stereotypes computed from two items from the PISA survey, which are the gender gap in attribution to failure in math and the gender gap in mathematics self-efficacy in calculating the petrol-consumption rate of a car (see data section for details). As displayed in Table S3, these measures are strongly correlated with the GMS.

Table S7A and S7B show that our results still hold with these alternative measures, except for our measure of "human equality" for which results are a bit less striking. Correlations with measures of development and inequality are significant and positive for both attribution to failure and self-efficacy, and of roughly the same magnitude as the GMS. The horse race regressions show that the measures of development and inequality lose all explanatory power once we introduce our two alternative measures of gender-math stereotypes (except for the measure of "human equality" whose effect is only divided by two and remains significant). The estimated effect of our alternative measures of gender math-related stereotypes on the gender gap is also in line with the estimates found for GMS in Table 1.

Estimates from individual-level linear regressions of the relation between GMS and countries development and equality (Table S8)

To show that our results are not driven by unobserved heterogeneity across students from more or less developed countries (*e.g.* differences in their socioeconomic backgrounds, their resources at home or the characteristics of the educational systems in which they study), we estimate individual-level linear regressions of the relationship between Gender-Math Stereotypes and countries' development and inequality (see the methods section for details). This is also a way to provide an alternative statistical inference for our main results (see methods section). Table S8 shows estimations where the standardized arithmetic average between B1 and B2 is regressed on a dummy equal to one for female students, and on the interaction terms between gender and our five measures of development. These five regressions are the micro equivalent of the results displayed in Panel A of Table 1.

We consider four different specifications in Table S8: column 1 only includes country fixed effects, column 2 includes country fixed effects and a control for students' math ability, column 3 adds a control for students' stated interest in math (in the spirit of our macro-level measure GMS2), column 4 includes an additional set of individual controls that could be confounding factors, namely the level of education of the student's parents (in number of years and degree obtained), grade repetition, an index of economic, social and cultural status of the household, a measure of home educational resources, and a measure of attitude towards school (how much students think that trying hard at school is important). Column 2 of Table S8 A is the closest equivalent to Panel

A of Table 1. We can see that gender, and all measures of development and equality interacted with gender are positively associated with GMS, showing that gender math-related stereotypes are larger in high income and more equal countries. This relationship holds and is statistically significant in all specifications, even once we add all individual controls, confirming that results in Panel A of Table 1 are not driven by the most obvious sources of unobserved heterogeneity at the student level.

Relationship between GMS and intentions to study math: estimates from individual-level linear regressions (Table S9)

Table S9 shows results for individual-level linear regressions of intentions to study math on gender math-related stereotypes. Each column of the table presents a different specification: the first one with no control, the second one controlling for math ability, the third controlling for math ability and math preferences, and finally the fourth column adding other individual controls.

Table S9 shows that the interaction term between the dummy for being a female student and the GMS is significantly different from zero and positive in all specifications, confirming that the gender gap in intentions to study math is larger for students in countries where gender-math stereotypes are stronger. This is true also when students' heterogeneity in terms of *e.g.* socioeconomic background and home resources is controlled for. The estimated effect varies from 0.10 to 0.05 in the more demanding specification. However, we lose about half of our observations for this last model, which also implies that we lose precision.

Note that we have examined if results in Table S9 are driven by girls being less willing to study math in countries with stronger stereotypes, or by boys being more willing to study math in these countries. To do so, we have removed the country fixed-effects from the empirical models and have added instead an interaction between GMS and a dummy for boys. Results (not reported) show no clear pattern that a gender rather than the other one drives the relationship between GMS and the gender gap in intentions to study math. However, the necessary absence of country-fixed in the corresponding specifications implies that this latter conclusion should be considered with caution.

Evidence of the gender equality paradox at the individual level (Table S10)

Table S10 shows results of individual-level linear regression models with intentions to study math as the dependent variable, and gender and an interaction term between gender and a measure of countries' development or extent of equality as regressors. The purpose of this table is to investigate the existence of the "Gender Equality Paradox" (*3*) at the micro level. Each column of Table S10 shows a different specification in the same fashion as Table S9: estimations with country fixed effects only are presented in column 1, estimations with additional controls for math ability, preferences and individual constant characteristics are displayed in column 2, 3 and 4.

Except for GDP, the estimated effect of the interaction between development or equality and gender is negative and significantly different from zero at the 1% or 5% level in each of the four tested specifications. For the concerned variables, this confirms that the gender gap (boys minus girls) in intentions to study math grows with development or more equality even when controlling for students' observable characteristics. This is not the case however when economic development is considered: the relationship between GDP and the gender gap in intentions to study math becomes weaker and not statistically significant once control variables for students' characteristics are included. Comparing columns 2 and 3, we see that this is driven by the control for students' interest for math, while other controls have limited effect on their own on the relationship between GDP and the gender gap in intentions to study math.

Stereotypes mediate the "Gender Equality Paradox": individual-level evidence (Table S11)

Table S11 complements Table 2 and displays the results for the individual-level counterpart of the "horse race" regressions provided in Panel B of Table 1. Intentions to study math are regressed on a dummy for female students, on an interaction term between gender and GMS, and an interaction term between gender and a measure of development or equality. As for previous tables, estimations with country fixed effects are presented in column 1, estimations with additional controls for math ability, preferences and individual constant characteristics are displayed in column 2, 3 and 4.

With a few exceptions, the micro linear models lead to the same conclusions as the macro models presented in Table 1. In all specifications, the estimated effect of variables capturing development or equality is largely reduced once we introduce the interaction between the dummy for being a female student and the GMS (this can be seen by comparing the estimates in Tables S10 and S11) and it becomes statistically non-significant in most cases. In contrast, the estimated effect of the interaction term girl*GMS is not affected much by the inclusion of an interaction between gender and countries' development. As can be seen in the fourth column of Table 2 and Table 11 (first panel), the relationship between GDP and the gender gap un intentions is even reversed once students' characteristics and GMS interacted with gender are controlled for. The estimate of 0.0259 indeed implies that relative to boys, girls intend more to pursue math-related studies and careers in more developed countries

Relationship between gender math-related and other types of gender stereotypes or cultural norms (Table S12)

Table S12 displays the correlation matrix between measures of gender math-related stereotypes and measures of other types of gender stereotypes or countries' cultural norms.

Liberal values and essentialist gender norms

In the main text, we argue that "the reason why gender essentialist norms (regarding math or other domains) are more pronounced in more developed and egalitarian countries could be that countries with more emancipative, individualistic and progressive values "give their citizens greater space to fall back on an old, deeply ingrained cultural frame as they try to make sense of themselves and others and organize their choices and behaviors accordingly" *(14)*. Gender is likely to be among the most important of those "deeply ingrained cultural frames" used to organize social relations

(14,15). Hence, the fact that more developed and egalitarian countries may give their citizens more space to fall back on such cultural frames could explain that the interiorization of the old stereotype that math is not for girls" (captured by GMS) is stronger in these countries.

To back up partially this claim, we show in Table S12 that gender-math stereotypes captured by GMS are indeed stronger in countries having more individualistic or emancipative values (which are also more developed and more equal on average). The correlations of GMS with the two indexes capturing this type of values are large (r=0.61 and r=0.62) and highly significant (p<0.01).

Gender-math stereotypes and other essentialist gender norms

Table S12 also shows that gender-math stereotypes captured with GMS are strongly correlated at the country-level with other types of essentialist gender norms or attributes that are also more widespread in more developed countries. They are for example strongly correlated with gender gaps in self-reported personality traits such as feminine openness (r=0.70) or feminine extraversion (r=0.76) or in the gender gap in the importance granted to the family (r=0.74) which can capture how both genders interiorize differentially their reserved domains (e.g., math for boys, family or care for girls). We also find a positive correlation between GMS and measures of gender gaps in emotions, such as the gender gap in crying tendency (results not reported). These results show that gender math stereotypes can be linked to other types of gender norms or gender differences that tend to develop more in wealthier or more egalitarian countries. This suggests that our main conclusions that are specific to the gender equality paradox regarding occupational segregation may extend to gender equality paradoxes regarding other gender gaps. More precisely, the fact that several gender gaps are larger in wealthier or more egalitarian countries could be due to the fact that these countries that have eliminated at least to some extent the male primacy ideology have developed other types of (more horizontal) gender essentialist norms.

Horizontal and vertical stereotypes

Table S12 (and Figure 2) finally shows the negative correlations between the "horizontal" essentialist gender attributes discussed above and more "vertical" gender stereotypes that can be directly linked to traditional gender roles and the male primacy ideology. All of the 16 pairwise correlations between the four measures of "horizontal" stereotypes (GMS and gender gaps in feminine extraversion, feminine openness and the importance granted to the family) and the four measures of the importance of traditional gender roles ("Women should take care of their home and children", "Being a housewife is fulfilling", "University is more important for a boy than a girl" and the index capturing negative values regarding gender equality in in education, politics and on the labor market) are negative, 12 of them being statistically significant at the 5% level. In particular, the correlation of GMS is negative and significant at the 5% level with the four measures of vertical stereotypes. These simple correlations provide evidence supporting the claim that "the countries that have eliminated at least to some extent the male primacy ideology and traditional gender roles have developed other types of (more horizontal) gender essentialist norms". An interesting avenue for future research would be to study the joint evolution over time of these different types of gender norms. This would make it possible to trace more precisely the gender norms that develop and those that disappear when countries develop or become more egalitarian or individualistic. Unfortunately, we have not been able to collect enough data on stereotypes and gender norms back in time to perform such an exercise convincingly (see nevertheless next section for some limited analyses of the joint evolution of GDP or inequality and GMS).

Average Performance in Mathematics

More developed and egalitarian countries tend to have higher levels of math performance, that are likely to be associated with higher internalized gender math stereotypes. We find that GMS is positively correlated with countries' average performance in math, and that average performance in math is positively correlated with development as measured by the GDP. This result is in line with recent research in sociology (*16*), which has shown that stronger country performance implies more difficult curricula, higher performance standards, and greater competition, all of which heighten gender essentialist ideas about math and science.

Dynamic relationships between measures of development and gender math-related stereotypes (Table S13)

Table S13 displays the correlation matrix between GDP growth or changes in the opposite of the GINI coefficient and changes in the four gender gaps we use as "the best available proxies" for Gender-Math stereotypes over the period 2003-2012. We do not include the GGI, the Human Development Index and the Coefficient of Human Inequality in these analyses because those measures are not available in 2003.[5] The four proxies we use are the gender gaps in four questions capturing intrinsic motivation for math that are included in both PISA 2003 and PISA 2012. The gender gaps for intrinsic motivation are available for 40 countries of our original sample, among which we have value for GDP growth for 39 countries and values for the opposite of the GINI for 32 countries.

We find no clear relationships between changes in our proxies for gender-math stereotypes and changes in inequality captured by the GINI index: all estimated correlations are small and not significantly different from zero. The small number of countries kept for the analysis and the limited time span over which the evolutions in inequality or stereotypes are considered may explain that we are not able to detect any relationship.

In contrast, we find a positive dynamic association between each of the four proxies of gender-math stereotypes we use and economic development: on average, gender-math stereotypes increased relatively more (or decreased relatively less) in countries that experienced a larger economic growth over the period 2003-2012. The correlation is significant at the 5% level for one proxy and at the 10% level for another one. This provides weak evidence that economic development can go hand in hand with an increase in the importance of gender essentialist norms. At least, Table S13 makes clear that economic development has not led to a reduction of gender essentialist norms such as gender-math stereotypes over the first decade of the 21st century.

[5] The Human Development Index does exist in 2003 but it is based on an alternative definition so that it cannot be directly compared with 2012.

References

1. OECD, PISA 2012 Results (Volume I): Excellence and Equity in Education, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264266490-en (2016)
2. OECD, PISA 2012 Technical Report. OECD Publishing, Paris. http://www.oecd.org/pisa/data/2015-technical-report/
3. G. Stoet, D.C. Geary, The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, **29**(4), 581–593. (2018)
4. G. Stoet, D.C. Geary "The gender-equality paradox in science, technology, engineering, and mathematics education": Corrigendum. (2020)
5. S. S. Richardson, M.V. Reiches, J. Bruch, M. Boulicault, N.E. Noll, & H. Shattuck-Heidorn, Is there a gender-equality paradox in science, technology, engineering, and math (STEM)? Commentary on the study by Stoet and Geary (2018). *Psychological Science*, *31*(3), 338-341. (2020).
6. M. Charles, K. Bradley, Indulging our gendered selves? Sex segregation by field of study in 44 countries. *American journal of sociology*, 114(4), 924-976. (2009)
7. D.I. Miller, A.H. Eagly, M.C. Linn, Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, 107(3), 631-644. (2015). https://doi.org/10.1037/edu0000005
8. C.R. Knight, M.C. Brinton. One egalitarianism or several? Two decades of gender-role attitude Change in Europe. AJS Volume 122 Number 5 (March 2017): 1485–1532
9. P.T. Costa Jr, A. Terracciano, R.R. McCrae, Gender differences in personality traits across cultures: robust and surprising findings. Journal of personality and social psychology, 81(2), 322. (2001)
10. J.M. Wooldridge, Introductory econometrics: A modern approach. Nelson Education (2016)
11. M.E. Sobel, Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*. **13**: 290–312. (1982).
12. M.E. Sobel, Some New Results on Indirect Effects and Their Standard Errors in Covariance Structure". *Sociological Methodology.* **16:** 159–186. (1986)
13. W.H. Yu, P.L. Lee, Decomposing gender beliefs: Cross-national differences in attitudes toward maternal employment and gender equality at home. *Sociological Inquiry*, *83*(4), 591-621. (2013)
14. C.L. Ridgeway, Framed before we know it: How gender shapes social relations. *Gender & society*, 23(2), 145-160. (2009)
15. C.L. Ridgeway, Gender as a group process: Implications for the persistence of inequality. In The social psychology of gender, ed. S. J. Correll. New York: Elsevier (2007)
16. A. Mann, T.A. DiPrete, The consequences of the national math and science performance environment for gender differences in STEM aspiration. *Sociological Science*, *3*, 568-603.(2016)

Figures and Tables

Table S1: Marginal effects of math performance, coefficients from the GMS regressions

	Marginal effect of math performance on students' answers to		
	B1: "Doing well in math is completely up to me"	**B2: "My parents believe that math is important for my career"**	**(B1+B2)/2**
Among OECD countries			
Australia	-0.159***	-0.082***	-0.155***
Austria	0.051**	0.139***	0.123***
Belgium	0.035**	-0.070***	-0.018
Canada	-0.134***	-0.132***	-0.172***
Switzerland	0.029	0.066***	0.062***
Chile	-0.032**	-0.026	-0.035**
Czech Republic	0.024	0.004	0.018
Germany	0.064***	0.128***	0.124***
Denmark	-0.178***	-0.065***	-0.155***
Spain	-0.012	-0.127***	-0.088***
Estonia	-0.107***	-0.041**	-0.094***
Finland	-0.162***	-0.167***	-0.206***
France	0.007	-0.085***	-0.050**
United Kingdom	-0.090***	0.016	-0.051**
Greece	-0.045**	-0.101***	-0.095***
Hungary	0.013	0.024	0.023
Ireland	-0.089***	-0.024	-0.073***
Iceland	-0.237***	-0.150***	-0.241***
Israel	0.031	-0.047**	-0.012
Italy	0.002	-0.033***	-0.019*
Japan	-0.151***	-0.203***	-0.230***
Korea	-0.320***	-0.326***	-0.402***
Luxembourg	-0.080***	-0.013	-0.062***
Mexico	-0.142***	-0.007	-0.092***
Netherlands	-0.102***	-0.061**	-0.101***
Norway	-0.218***	-0.171***	-0.247***
New Zealand	-0.140***	-0.092***	-0.150***
Poland	-0.094***	-0.125***	-0.144***
Portugal	-0.035	-0.203***	-0.153***
Slovak Republic	-0.008	0.015	0.006
Slovenia	-0.109***	-0.022	-0.085***
Sweden	-0.109***	-0.031	-0.092***
Turkey	-0.123***	-0.131***	-0.163***
United States	-0.143***	-0.071***	-0.139***
Among non-OECD countries			
Albania	0.012	0.025	0.021
United Arab Emirates	-0.054***	-0.087***	-0.088***
Argentina	-0.065***	0.063***	-0.002
Bulgaria	-0.065***	-0.011	-0.047**
Brazil	-0.013	0.039***	0.012
Colombia	-0.072***	0.051**	-0.014

Costa Rica	0.061***	0.035	0.060***
Hong Kong	-0.113***	-0.121***	-0.150***
Croatia	-0.007	-0.012	-0.012
Indonesia	-0.159***	0.014	-0.092***
Jordan	-0.032	-0.109***	-0.090***
Kazakhstan	-0.182***	-0.023	-0.130***
Liechtenstein	0.159**	0.269***	0.263***
Lithuania	-0.105***	-0.073***	-0.114***
Latvia	-0.092***	-0.088***	-0.111***
Macao	-0.125***	-0.017	-0.093***
Montenegro	-0.061***	0.013	-0.031
Malaysia	-0.144***	-0.079***	-0.143***
Peru	-0.143***	-0.050**	-0.121***
Qatar	-0.105***	-0.177***	-0.174***
Shanghai-China	-0.097***	-0.022	-0.076***
Romania	-0.044*	-0.019	-0.040*
Russia	-0.154***	0.001	-0.095***
Singapore	-0.121***	0.111***	-0.007
Serbia	-0.020	0.010	-0.006
Chinese Tapeï	-0.257***	-0.164***	-0.270***
Thailand	0.029*	-0.020	0.007
Tunisia	-0.066***	-0.163***	-0.148***
Uruguay	-0.016	0.053***	0.023
Viet Nam	-0.036*	-0.141***	-0.116***

Notes: Regressions are done country by country, and systematically include a dummy for girls whose estimated effect is given in Table S2. All variables are standardized to have a weighted mean equal to 0 and a weighted standard deviation equal to 1 in each country. Estimates and standard errors involving measures of ability are based on plausible values and account for measurement error in these abilities on top of standard sampling error. See section "Country-Level Variables Computed from PISA Items" of this SI for more details. *** *p<0.01*, ** *p<0.05*, * *p<0.1*

Table S2 : Measures of gender math-related stereotypes: averages and breakdown by country

	Measures of gender math-related stereotypes				
	B1: "Doing well in math is completely up to me"		**B2: "My parents believe that math is important for my career"**		**GMS**
	Raw gender gap	Gender gap conditional on math performance	Raw gender gap	Gender gap conditional on math performance	
Average: over all countries in the sample	0.10	0.10	0.15	0.15	0.16
Average: over OECD countries	0.14	0.14	0.2	0.19	0.21
Average: over Non-OECD countries	0.05	0.05	0.1	0.11	0.1
Among OECD countries					
Australia	0.177***	0.155***	0.291***	0.279***	0.278***
Austria	0.178***	0.190***	0.344***	0.378***	0.365***
Belgium	0.179***	0.183***	0.356***	0.347***	0.346***
Canada	0.152***	0.134***	0.119***	0.101***	0.148***
Switzerland	0.226***	0.231***	0.387***	0.396***	0.400***
Chile	0.087**	0.077**	0.200***	0.193***	0.173***
Czech Republic	0.184***	0.189***	0.319***	0.320***	0.328***
Germany	0.317***	0.327***	0.294***	0.314***	0.411***
Denmark	0.277***	0.242***	0.234***	0.221***	0.291***
Spain	0.104***	0.101***	0.156***	0.125***	0.149***
Estonia	0.072*	0.062	0.210***	0.206***	0.172***
Finland	0.210***	0.205***	0.139***	0.136***	0.217***
France	0.195***	0.196***	0.245***	0.234***	0.281***
United Kingdom	0.123***	0.113***	0.194***	0.196***	0.206***
Greece	0.117***	0.112***	0.184***	0.173***	0.180***
Hungary	0.251***	0.252***	0.359***	0.362***	0.380***
Ireland	0.150***	0.131***	0.218***	0.213***	0.218***
Iceland	0.087*	0.100**	0.108**	0.116***	0.133***
Israel	0.141***	0.148***	0.217***	0.207***	0.237***
Italy	0.139***	0.139***	0.149***	0.142***	0.186***
Japan	-0.006	-0.040	0.152***	0.108***	0.045
Korea	0.091**	0.026	0.214***	0.148***	0.108***
Luxembourg	0.167***	0.144***	0.235***	0.231***	0.234***
Mexico	-0.044**	-0.067***	0.092***	0.091***	0.017
Netherlands	0.333***	0.317***	0.255***	0.246***	0.357***
Norway	0.155***	0.151***	0.055	0.056*	0.127***
New Zealand	0.109***	0.084**	0.204***	0.187***	0.178***
Poland	0.162***	0.155***	0.088**	0.078**	0.149***
Portugal	0.047	0.041	0.010	-0.020	0.010

Slovak Republic	0.142***	0.141***	0.258***	0.260***	0.257***
Slovenia	0.058	0.050	0.218***	0.217***	0.165***
Sweden	0.200***	0.203***	0.178***	0.178***	0.244***
Turkey	0.033	0.022	-0.039	-0.050	-0.015
United States	0.093***	0.077**	0.062*	0.054*	0.082**
Among non-OECD countries					
Albania	0.010	0.010	0.067*	0.066*	0.053
United Arab Emirates	-0.048	-0.047	0.143***	0.145***	0.066*
Argentina	-0.026	-0.040	0.038	0.052	0.008
Bulgaria	-0.003	-0.003	0.154***	0.154***	0.103***
Brazil	0.029	0.026	0.082***	0.089***	0.073***
Colombia	-0.074*	-0.096**	0.065*	0.081**	-0.016
Costa Rica	0.107***	0.129***	0.175***	0.187***	0.203***
Hong Kong	0.270***	0.250***	0.105***	0.084**	0.217***
Croatia	0.154***	0.153***	0.228***	0.226***	0.241***
Indonesia	-0.069*	-0.082**	0.018	0.018	-0.043
Jordan	0.022	0.027	0.069*	0.088***	0.077**
Kazakhstan	-0.133***	-0.133***	0.029	0.029	-0.063*
Liechtenstein	0.165	0.231	0.332**	0.445***	0.451***
Lithuania	0.031	0.028	0.274***	0.272***	0.189***
Latvia	0.068	0.075*	0.187***	0.192***	0.170***
Macao	0.294***	0.293***	0.073**	0.073**	0.238***
Montenegro	0.016	0.016	0.168***	0.168***	0.118***
Malaysia	-0.066*	-0.057	-0.095***	-0.089**	-0.091***
Peru	-0.041	-0.071*	-0.025	-0.036	-0.060
Qatar	0.010	0.019	0.087***	0.103***	0.075***
Shanghai-China	0.206***	0.200***	-0.004	-0.005	0.124***
Romania	0.014	0.012	0.009	0.008	0.016
Russia	0.058*	0.061*	0.292***	0.292***	0.217***
Singapore	0.056	0.058*	0.156***	0.154***	0.138***
Serbia	0.162***	0.159***	0.256***	0.257***	0.271***
Chinese Tapeï	0.132***	0.113***	0.164***	0.151***	0.168***
Thailand	-0.009	-0.013	-0.075**	-0.072**	-0.053
Tunisia	0.041	0.024	0.049	0.007	0.029
Uruguay	0.111***	0.109***	0.168***	0.175***	0.196***
Viet Nam	0.085**	0.081**	-0.042	-0.058*	0.014

Notes: Regressions are done country by country, and systematically include a variable accounting for math performance whose estimated effect is given in Table S1. All variables are standardized to have a weighted mean equal to 0 and a weighted standard deviation equal to 1 in each country. Estimates and standard errors involving measures of ability are based on plausible values and account for measurement error in these abilities on top of standard sampling error. See section "Country-Level Variables Computed from PISA Items" of this SI for more details. *** *p<0.01*, ** *p<0.05*, * *p<0.1*

Table S3: Correlation matrix between GMS and alternative measures of gender-math stereotypes

	GMS and alternative measures of gender math-related stereotypes					
	GMS	**Gender gap in B1 (conditional on math ability)**	**Gender gap in B2 (conditional on math ability)**	**Gender gap in mathematics self-efficacy - petrol consumption rate**	**Gender gap in attribution to failure in mathematics**	**IAT**
GMS: main measure of Gender-Math Stereotypes	1					
Gender gap in B1 (conditional on math ability)	0.872***	1				
Gender gap in B2 (conditional on math ability)	0.903***	0.578***	1			
Gender gap in math self-efficacy (confidence in own ability to calculate the petrol consumption rate of a car)	0.685***	0.519***	0.692***	1		
Gender gap in attribution to failure in mathematics	0.735***	0.722***	0.596***	0.539***	1	
Implicit Association Test relating science to men	0.640***	0.590***	0.573***	0.605***	0.392***	1

Notes: Details for the computations of GMS, gender gaps in B1, B2, mathematics self-efficacy - petrol consumption rate, attribution to failure in mathematics, and source for the IAT measure are available in the SI Data section.

Table S4A: Relations between Gender-math stereotypes, women's underrepresentation in math-related fields and countries' development or inequality for OECD countries

Panel A) Correlation of the Gender-Math Stereotype (GMS) with measures X of countries' development and economic, human or gender equality

	Measure X of countries' development or equality				
	GDP	**Human Development Index**	**Income equality (opposite of GINI index)**	**"Human equality" (opposite of coefficient of Human Inequality)**	**Gender equality (Gender Gap Index)**
Correlation of GMS with…	0.293*	0.480***	0.470***	0.639***	0.291*

Panel B) Comparing the explanatory power of countries' Gender-Math Stereotypes (GMS) versus countries' development or equality measures X to predict gender gaps in students' intentions to pursue math-related studies or careers (Y). Estimates from linear regression models with one (univariate) or two (horse race) regressors

	Y=Gender gap (boys minus girls) in intentions to study math or pursue a math career (standardized index)				
Marginal effect of GMS on Y (regression of Y on GMS when X is available)	0.592***	0.592***	0.592***	0.604***	0.592***
Marginal effect of X on Y (regression of Y on X)	0.205	0.355**	0.275	0.560***	0.275
Marginal effect of X on Y controlling for GMS (regression of Y on X and GMS)	0.0350	0.0925	-0.00485	0.294	0.113
Marginal effect of GMS on Y controlling for X (regressions of Y on X and GMS)	0.582***	0.548***	0.595***	0.416**	0.560***
Number of countries	34	34	34	33	34

Notes: All variables are standardized before each correlation or regression to have a mean equal to 0 and a standard deviation equal to 1 on the sample of countries included in the analysis. See the data section in this SI for details on the construction of GMS and the sources of country-level measures of development or equality. *** *p<0.01*, ** *p<0.05*, * *p<0.1*.

Table S4B: Relations between Gender-math stereotypes, women's underrepresentation in math-related fields and countries' development or inequality for non-OECD countries

Panel A) Correlation of the Gender-Math Stereotype (GMS) with measures X of countries' development and economic, human or gender equality

	Measure X of countries' development or equality				
	GDP	**Human Development Index**	**Income equality (opposite of GINI index)**	**"Human equality" opposite of coefficient of Human Inequality)**	**Gender equality (Gender Gap Index)**
Correlation of GMS with…	0.323	0.593***	0.0912	0.484**	0.336

Panel B) Comparing the explanatory power of countries' Gender-Math Stereotypes (GMS) versus countries' development or equality measures X to predict gender gaps in students' intentions to pursue math-related studies or careers (Y). Estimates from linear regression models with one (univariate) or two (horse race) regressors

	Y=Gender gap (boys minus girls) in intentions to study math or pursue a math career (standardized index)				
Marginal effect of GMS on Y (regression of Y on GMS when X is available)	0.627***	0.772***	0.750***	0.714***	0.702***
Marginal effect of X on Y (regression of Y on X)	0.0487	0.420**	-0.0381	0.307	0.347*
Marginal effect of X on Y controlling for GMS (regression of Y on X and GMS)	-0.171	-0.0587	-0.107	-0.0494	0.126
Marginal effect of GMS on Y controlling for X (regressions of Y on X and GMS)	0.682***	0.807***	0.760***	0.738***	0.659***
Number of countries	27	27	22	21	25

Notes: All variables are standardized before each correlation or regression to have a mean equal to 0 and a standard deviation equal to 1 on the sample of countries included in the analysis. See the data section in this SI for details on the construction of GMS and the sources of country-level measures of development or equality. *** *p<0.01*, ** *p<0.05*, * *p<0.1.*

Table S5: Relations between Gender-math stereotypes, women's underrepresentation in math-related fields (occupational segregation) and countries' development or inequality

Comparing the explanatory power of countries' Gender-Math Stereotypes (GMS) versus countries' development or equality measures X to predict alternative outcomes of occupational segregation (Y). Estimates from linear regression models with one (univariate) or two (horse race) regressors

	Measure X of countries' development or equality				
	GDP	**Human Development Index**	**Income equality (opposite of GINI index)**	**"Human equality" (opposite of coefficient of Human Inequality)**	**Gender equality (Gender Gap Index)**
	Y=Measure of segregation in science (*(3)*, standardized index)				
Marginal effect of GMS on Y (regression of Y on GMS when X is available)	-0.561***	-0.561***	-0.571***	-0.571***	-0.561***
Marginal effect of X on Y (regression of Y on X)	-0.318**	-0.498***	-0.143	-0.381**	-0.519***
Marginal effect of X on Y controlling for GMS (regression of Y on X and GMS)	-0.0603	-0.196	0.241	0.0398	-0.321**
Marginal effect of GMS on Y controlling for X (regressions of Y on X and GMS)	-0.532***	-0.420**	-0.703***	-0.599***	-0.403***
Number of countries	45	45	43	42	45
	Y=Female underrepresentation in Science in tertiary education (*(7)*, standardized index)				
Marginal effect of GMS on Y (regression of Y on GMS when X is available)	-0.597***	-0.643***	-0.622***	-0.600***	-0.606***
Marginal effect of X on Y (regression of Y on X)	-0.403***	-0.559***	-0.211	-0.366***	-0.329**
Marginal effect of X on Y controlling for GMS (regression of Y on X and GMS)	-0.176	-0.266**	0.0835	0.0555	-0.102
Marginal effect of GMS on Y controlling for X (regressions of Y on X and GMS)	-0.521***	-0.481***	-0.659***	-0.637***	-0.565***
Number of countries	55	55	51	49	53
	Y=Female overrepresentation in humanities (*(6)*, standardized index)				
Marginal effect of GMS on Y (regression of Y on GMS when X is available)	0.461***	0.461***	0.474***	0.506***	0.474***
Marginal effect of X on Y (regression of Y on X)	0.208	0.405**	0.297*	0.513***	0.272*
Marginal effect of X on Y controlling for GMS (regression of Y on X and GMS)	-0.0356	0.182	0.0351	0.309	0.0600
Marginal effect of GMS on Y controlling for X (regressions of Y on X and GMS)	0.479***	0.342*	0.454**	0.285	0.446**
Number of countries	39	39	38	36	38

Notes: All variables are standardized before each correlation or regression to have a mean equal to 0 and a standard deviation equal to 1 on the sample of countries included in the analysis. See the data section in this SI for details on the construction of GMS and the sources of country-level measures of development, equality or gender segregation across fields. *** *p<0.01*, ** *p<0.05*, * *p<0.1*.

Table S6: Relations between Gender-math stereotypes (controlling for students' interest for math), women's underrepresentation in math-related fields and countries' development or inequality

Panel A) correlation of Gender-Math Stereotypes conditional on both interest for math and math ability (GMS2) with measures X of countries' development and economic, human or gender equality

	Measure X of countries' development or equality				
	GDP	**Human Development Index**	**Income equality (opposite of GINI index)**	**"Human equality" (opposite of coefficient of Human Inequality)**	**Gender equality (Gender Gap Index)**
Correlation of GMS2 with…	0.353***	0.613***	0.514***	0.708***	0.514***

Panel B) Comparing the explanatory power of countries' GMS2 versus countries' development or equality measures X to predict gender gaps in students' intentions to pursue math-related studies or careers (Y). Estimates from linear regression models with one (univariate) or two (horse race) regressors

	Y=Gender gap (boys minus girls) in intentions to study math or pursue a math career (standardized index)				
Marginal effect of GMS2 on Y (regression of Y on GMS2 when X is available)	0.581***	0.612***	0.608***	0.598***	0.589***
Marginal effect of X on Y (regression of Y on X)	0.326**	0.498***	0.300**	0.545***	0.402***
Marginal effect of X on Y controlling for GMS2 (regression of Y on X and GMS2)	0.138	0.197	-0.0175	0.244	0.177
Marginal effect of GMS2 on Y controlling for X (regressions of Y on X and GMS2)	0.533***	0.492***	0.617***	0.425***	0.511***
Number of countries	61	61	56	54	59

Notes: All variables are standardized before each correlation or regression to have a mean equal to 0 and a standard deviation equal to 1 on the sample of countries included in the analysis. See the data section in this SI for details on the construction of GMS2 and the sources of country-level measures of development or equality. *** *p<0.01*, ** *p<0.05*, * *p<0.1*.

Table S7A: Relations between gender differences in attribution to failure in mathematics, women's underrepresentation in math-related fields and countries' development or inequality

Panel A) Correlation of gender differences in attribution to failure in mathematics (GMS3) with measures X of countries' development and economic, human or gender equality

	Measure X of countries' development or equality				
	GDP	**Human Development Index**	**Income equality (opposite of GINI index)**	**"Human equality" (opposite of coefficient of Human Inequality)**	**Gender equality (Gender Gap Index)**
Correlation of GMS3 with…	0.558***	0.599***	0.464***	0.685***	0.648***

Panel B) Comparing the explanatory power of countries' alternative GMS3 versus countries' development or equality measures X to predict gender gaps in students' intentions to pursue math-related studies or careers (Y). Estimates from linear regression models with one (univariate) or two (horse race) regressors

	Y=Gender gap (boys minus girls) in intentions to study math or pursue a math career (standardized index)				
Marginal effect of GMS3 on Y (regression of Y on GMS3 when X is available)	0.578***	0.606***	0.595***	0.579***	0.583***
Marginal effect of X on Y (regression of Y on X)	0.326**	0.498***	0.300**	0.545***	0.402***
Marginal effect of X on Y controlling for GMS3 (regression of Y on X and GMS3)	0.00495	0.210	0.0297	0.279*	0.0418
Marginal effect of GMS3 on Y controlling for X (regressions of Y on X and GMS3)	0.575***	0.481***	0.582***	0.388**	0.556***
Number of countries	61	61	56	54	59

Notes: All variables are standardized before each correlation or regression to have a mean equal to 0 and a standard deviation equal to 1 on the sample of countries included in the analysis. See the data section in this SI for details on the construction of GMS3 and the sources of country-level measures of development or equality. *** *p<0.01*, ** *p<0.05*, * *p<0.1*.

Table S7B: Relations between gender differences in mathematics self-efficacy in calculating petrol consumption rate, women's underrepresentation in math-related fields and countries' development or inequality

Panel A) Correlation of gender differences in mathematics self-efficacy in calculating petrol consumption rate (GMS4) with measures X of countries' development and economic, human or gender equality

	GDP	Human Development Index	Income equality (opposite of GINI index)	"Human equality" (opposite of coefficient of Human Inequality)	Gender equality (Gender Gap Index)
Correlation of GMS4 with…	0.326**	0.602***	0.482***	0.622***	0.513***

Panel B) Comparing the explanatory power of countries' GMS4 versus countries' development or equality measures X to predict gender gaps in students' intentions to pursue math-related studies or careers (Y). Estimates from linear regression models with one (univariate) or two (horse race) regressors

	Y=Gender gap (boys minus girls) in intentions to study math or pursue a math career (standardized index)				
Marginal effect of GMS4 on Y (regression of Y on GMS4 when X is available)	0.603***	0.616***	0.592***	0.569***	0.598***
Marginal effect of X on Y (regression of Y on X)	0.326**	0.498***	0.300**	0.545***	0.402***
Marginal effect of X on Y controlling for GMS4 (regression of Y on X and GMS4)	0.145	0.199	0.0186	0.311**	0.129
Marginal effect of GMS4 on Y controlling for X (regressions of Y on X and GMS4)	0.555***	0.496***	0.583***	0.376**	0.531***
Number of countries	61	61	56	54	59

Notes: All variables are standardized before each correlation or regression to have a mean equal to 0 and a standard deviation equal to 1 on the sample of countries included in the analysis. See the data section in this SI for details on the construction of GMS4 and the sources of country-level measures of development or equality. *** *p<0.01*, ** *p<0.05*, * *p<0.1*.

Table S8: Relationship between the Gender-Math stereotype and countries' development or inequality. Estimates from individual-level linear regressions

	Dependent variable is standardized average of B1 and B2			
a) Linking the Gender-math stereotype to GDP				
Girl	0.165***	0.155***	0.101***	0.158***
(s.e)	(0.0137)	(0.0140)	(0.0128)	(0.0119)
Girl*GDP	0.0553***	0.0542***	0.0324**	0.0328**
(s.e)	(0.0145)	(0.0142)	(0.0128)	(0.0126)
Number of observations	304,252	304,252	303,406	139,058
b) Linking the Gender-math stereotype to Human development (HDI)				
Girl	0.164***	0.154***	0.102***	0.155***
(s.e)	(0.0117)	(0.0123)	(0.0111)	(0.0100)
Girl*HDI	0.0827***	0.0806***	0.0621***	0.0576***
(s.e)	(0.0105)	(0.0102)	(0.00859)	(0.00908)
Number of observations	300,910	300,910	300,060	137,417
c) Linking the Gender-math stereotype to income equality (opposite of GINI)				
Girl	0.149***	0.138***	0.0902***	0.146***
(s.e)	(0.0148)	(0.0150)	(0.0126)	(0.0116)
Girl*(-GINI)	0.0653***	0.0675***	0.0589***	0.0609***
(s.e)	(0.0158)	(0.0150)	(0.0123)	(0.0114)
Number of observations	279,951	279,951	279,229	127,467
d) Linking the Gender-math stereotype to "human equality"				
Girl	0.155***	0.145***	0.0959***	0.150***
(s.e)	(0.0121)	(0.0123)	(0.0105)	(0.0105)
Girl*"Human equality"	0.0893***	0.0903***	0.0784***	0.0724***
(s.e)	(0.0117)	(0.0109)	(0.00904)	(0.00847)
Number of observations	273,831	273,831	273,128	126,257
e) Linking the Gender-math stereotype to Gender equality (Gender Gap Index)				
Girl	0.156***	0.146***	0.0959***	0.153***
(s.e)	(0.0145)	(0.0147)	(0.0128)	(0.0122)
Girl*GGI	0.0464***	0.0463***	0.0411***	0.0341***
(s.e)	(0.0134)	(0.0126)	(0.0102)	(0.0103)
Number of observations	297,662	297,662	296,818	135,866
Country fixed effects	Yes	Yes	Yes	Yes
Individual control for math ability	No	Yes	Yes	Yes
Individual control for math preferences	No	No	Yes	Yes
Other individual controls	No	No	No	Yes

Notes: Other individual controls include the level of education of the student's parents, measured both in years and kind of diploma obtained, grade repetition, an index of economic, social and cultural status of the household, a measure of home educational resources, and a measure of attitude towards school, namely how much students think that trying hard at school is important. See the data section in this SI for details about the sources of country-level measures of development or equality. Standard errors have been clustered at the country level. Regressions are weighted by "senate" weights which sum to one in each country. See the method section of this SI for details on the empirical models. *** p<0.01, ** p<0.05, * p<0.1.

Table S9: Students' intentions to study math or pursue a math career and the Gender-Math stereotype. Estimates from individual-level linear regressions

	Dependent variable is students' intentions to pursue math related studies or careers				
Linking intentions to study math to the Gender-math stereotype (includes country fixed effects)					
Girl	-0.262***	-0.247***	-0.183***	-0.193***	-0.197***
(s.e.)	(0.0153)	(0.0151)	(0.0147)	(0.0159)	(0.0147)
Girl*GMS	-0.101***	-0.0958***	-0.0705***	-0.0750***	-0.0501**
(s.e.)	(0.0144)	(0.0141)	(0.0130)	(0.0162)	(0.0207)
Number of observations	301,360	301,360	299,950	137,545	121,361
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Individual control for math ability	No	Yes	Yes	Yes	Yes
Individual control for math preferences	No	No	Yes	Yes	Yes
Other individual controls	No	No	No	Yes	Yes
Controls for five measures of countries' development and equality interacted with girl	No	No	No	No	Yes

Notes: Other individual controls include the level of education of the student's parents, measured both in years and kind of diploma obtained, grade repetition, an index of economic, social and cultural status of the household, a measure of home educational resources, and a measure of attitude towards school, namely how much students think that trying hard at school is important. See the data section in this SI for details about the sources of country-level measures of development or equality. Standard errors have been clustered at the country level. Regressions are weighted by "senate" weights which sum to one in each country. See the method section of this SI for details on the empirical models. *** p<0.01, ** p<0.05, * p<0.1.

Table S10: Evidence of the gender equality paradox at the micro level

	Dependent variable is standardized intentions to study math or pursue a math career			
a) Linking the Gender-math stereotype to GDP				
Girl	-0.278***	-0.262***	-0.198***	-0.209***
(s.e)	(0.0196)	(0.0189)	(0.0168)	(0.0176)
Girl*GDP	-0.0477***	-0.0455**	-0.0207	-0.0169
(s.e)	(0.0179)	(0.0174)	(0.0157)	(0.0192)
Number of observations	293,782	293,782	292,395	133,808
b) Linking the Gender-math stereotype to Human development (HDI)				
Girl	-0.283***	-0.267***	-0.203***	-0.219***
(s.e)	(0.0184)	(0.0177)	(0.0152)	(0.0161)
Girl*HDI	-0.0785***	-0.0749***	-0.0526***	-0.0582***
(s.e)	(0.0177)	(0.0171)	(0.0148)	(0.0171)
Number of observations	290,474	290,474	289,085	132,183
c) Linking the Gender-math stereotype to income equality (opposite of GINI)				
Girl	-0.273***	-0.254***	-0.196***	-0.205***
(s.e)	(0.0208)	(0.0202)	(0.0168)	(0.0176)
Girl*(-GINI)	-0.0537**	-0.0564***	-0.0453**	-0.0489**
(s.e)	(0.0214)	(0.0206)	(0.0174)	(0.0186)
Number of observations	270,026	270,026	268,781	122,514
d) Linking the Gender-math stereotype to "human equality"				
Girl	-0.275***	-0.255***	-0.197***	-0.203***
(s.e)	(0.0186)	(0.0179)	(0.0150)	(0.0155)
Girl*"Human equality"	-0.0929***	-0.0938***	-0.0793***	-0.0834***
(s.e)	(0.0219)	(0.0209)	(0.0185)	(0.0202)
Number of observations	264,058	264,058	262,834	121,361
e) Linking the Gender-math stereotype to Gender equality (Gender Gap Index)				
Girl	-0.275***	-0.259***	-0.199***	-0.209***
(s.e)	(0.0194)	(0.0188)	(0.0155)	(0.0163)
Girl*GGI	-0.0585***	-0.0578***	-0.0511***	-0.0462**
(s.e)	(0.0188)	(0.0181)	(0.0151)	(0.0177)
Number of observations	287,241	287,241	285,858	130,647
Country fixed effects	Yes	Yes	Yes	Yes
Control for math ability	No	Yes	Yes	Yes
Control for math preferences	No	No	Yes	Yes
Other individual controls	No	No	No	Yes

Notes: Other individual controls include the level of education of the student's parents, measured both in years and kind of diploma obtained, grade repetition, an index of economic, social and cultural status of the household, a measure of home educational resources, and a measure of attitude towards school, namely how much students think that trying hard at school is important. See the data section in this SI for details about the sources of country-level measures of development or equality. Standard errors have been clustered at the country level. Regressions are weighted by "senate" weights which sum to one in each country. See the method section of this SI for details on the empirical models. *** p<0.01, ** p<0.05, * p<0.1.

Table S11: Stereotypes mediate the "Gender Equality Paradox", micro evidence

	Dependent variable is standardized intentions to study math or pursue a math career			
Girl	-0.269***	-0.253***	-0.191***	-0.198***
(s.e)	(0.0152)	(0.0147)	(0.0138)	(0.0147)
Girl*GDP	-0.000134	-1.06e-05	0.0173	0.0259*
(s.e)	(0.0127)	(0.0127)	(0.0129)	(0.0136)
Girl*GMS	-0.101***	-0.0969***	-0.0811***	-0.0828***
(s.e)	(0.0157)	(0.0155)	(0.0147)	(0.0158)
Number of observations	293,782	293,782	292,395	133,808
Girl	-0.272***	-0.257***	-0.196***	-0.208***
(s.e)	(0.0147)	(0.0143)	(0.0131)	(0.0137)
Girl*HDI	-0.00812	-0.00840	-0.000498	-0.000339
(s.e)	(0.0153)	(0.0150)	(0.0149)	(0.0140)
Girl*GMS	-0.0995***	-0.0940***	-0.0739***	-0.0788***
(s.e)	(0.0156)	(0.0155)	(0.0151)	(0.0172)
Number of observations	290,474	290,474	289,085	132,183
Girl	-0.279***	-0.260***	-0.201***	-0.208***
(s.e)	(0.0144)	(0.0141)	(0.0124)	(0.0134)
Girl*(-GINI)	0.00601	-0.000496	-0.00326	-0.00938
(s.e)	(0.0156)	(0.0157)	(0.0143)	(0.0151)
Girl*GMS	-0.110***	-0.103***	-0.0775***	-0.0706***
(s.e)	(0.0167)	(0.0164)	(0.0152)	(0.0173)
	270,026	270,026	268,781	122,514
Girl	-0.277***	-0.258***	-0.199***	-0.205***
(s.e)	(0.0157)	(0.0151)	(0.0134)	(0.0141)
Girl*"Human equality"	-0.0239	-0.0313	-0.0367**	-0.0459**
(s.e)	(0.0197)	(0.0195)	(0.0180)	(0.0191)
Girl*GMS	-0.0883***	-0.0800***	-0.0546***	-0.0469**
(s.e)	(0.0198)	(0.0196)	(0.0180)	(0.0191)
Number of observations	264,058	264,058	262,834	121,361
Girl	-0.272***	-0.256***	-0.197***	-0.206***
(s.e)	(0.0146)	(0.0142)	(0.0123)	(0.0131)
Girl*GGI	-0.0176	-0.0191	-0.0227*	-0.0190
(s.e)	(0.0131)	(0.0128)	(0.0116)	(0.0141)
Girl*GMS	-0.0974***	-0.0922***	-0.0677***	-0.0658***
(s.e)	(0.0155)	(0.0151)	(0.0140)	(0.0159)
Number of observations	287,241	287,241	285,858	130,647
Country fixed effects	Yes	Yes	Yes	Yes
Control for math ability	No	Yes	Yes	Yes
Control for math preferences	No	No	Yes	Yes
Other individual controls	No	No	No	Yes

Notes: Other individual controls include the level of education of the student's parents, measured both in years and kind of diploma obtained, grade repetition, an index of economic, social and cultural status of the household, a measure of home educational resources, and a measure of attitude towards school, namely how much students think that trying hard at school is important. See the data section in this SI for details about the sources of country-level measures of development or equality. Standard errors have been clustered at the country level. Regressions are weighted by "senate" weights which sum to one in each country. See the method section of this SI for details on the empirical models. *** p<0.01, ** p<0.05, * p<0.1.

Table S12: Correlation matrix between measures of gender math-related stereotypes and general gender stereotypes

	Measures of gender math-related stereotypes and traditional gender stereotypes											
	GMS	**Feminine extraver-sion**	**Feminine openness**	**Gender gap: family important (WVS)**	**Women should take care of their home (ISSP)**	**Being a house-wife is full-filling (ISSP)**	**"University is more important for boys than girls" (WVS 2010-1014)**	**Equa-lity index (WVS 2005-2009)**	**Indivi-dualism**	**Eman-cipative values (WVS)**	**Average performance in Mathematics**	**GDP per capita**
***Gender-math stereotypes:* GMS**	1											
Essentialist gender attributes ("horizontal" stereotypes):												
Gender gap in feminine extraversion *(9)*	0.763***	1										
Gender gap in feminine openness *(9)*	0.698***	0.830***	1									
Gender gap in the share of individuals agreeing with the statement that family is important (WWS 2010-2014)	0.738***	0.773***	0.819***	1								
Traditional gender roles ("vertical" stereotypes): belief by women and men that…												
Women should take care of their home and children (ISSP 2012)	-0.386**	-0.269	-0.300	-0.461*	1							
Being a housewife is fulfilling (ISSP 2012)	-0.392**	-0.606**	-0.491	-0.723***	0.385**	1						
"University is more important for boys than girls" (WVS 2010-1014)	-0.532***	-0.591**	-0.475	-0.440**	0.751***	0.488*	1					
Equality index (WVS 2005-2009)	-0.559***	-0.872***	-0.800***	-0.597***	0.694***	0.766***	0.822***	1				
"Liberal values":												
Individualism *(7)*	0.611***	0.776***	0.793***	0.640***	-0.665***	-0.330	-0.660***	0.686***	1			

Emancipative values (WVS 2005-2009)	0.621***	0.741***	0.739***	0.673***	-0.825***	-0.577***	-0.720***	0.893***	0.725***	1		
Math performance:												
Average performance in Mathematics	0.527***	0.264	0.112	0.480***	-0.585***	0.109	-0.218	0.282*	0.479***	0.434***	1	
Development:												
GDP per capita	0.468***	0.385	0.205	0.319*	-0.791***	-0.193	-0.355*	0.667***	0.684***	0.830***	0.529***	1

Notes: The table shows correlation between our measure of Gender-Math Stereotypes (GMS) and several other measures of gender stereotypes. Details about this measure are displayed in SI. *** *p<0.01*, ** *p<0.05*, * *p<0.1*.

Table S13: correlation matrix between changes in gender math-related stereotypes and changes in countries' measures of development and equality (2003-2012)

	Δ Gender gap in "Making effort in math is worth-while"	**Δ Gender gap in "Mathematic will help me get a job"**	**Δ Gender gap in "Math is important for my future study"**	**Δ Gender gap in "Math is important for my career"**	**Δ in log GDP**	**Δ in Income equality (oppo-site of GINI index)**
Changes in motivation regarding mathematics (2003 - 2012)						
Δ Gender gap: "Making effort in math is worthwhile"	1					
Δ Gender gap: "Mathematic will help me get a job"	0.734***	1				
Δ Gender gap: "Math is important for my future study"	0.842***	0.780***	1			
Δ Gender gap: "Math is important for my career"	0.845***	0.709***	0.873***	1		
Changes in measures of countries' development and equality (2003-2012)						
Δ in Log GDP	0.241	0.190	0.362**	0.413***	1	
Δ in Income equality (opposite of GINI index)	-0.006	-0.183	-0.098	-0.052	0.235	1

Notes Changes in gender gaps (Δ) in motivation are computed as the difference in gender gaps controlling for math ability between 2012 and 2003, based on the PISA surveys Gender gaps are computed controlling for math performances, see SI for details about the chosen items. The change in log GDP represents the difference between the logarithm of GDP/capita for year 2012 and for year 2003, and the change in income equality the difference between the opposite of the GINI between 2012 and 2003. See the data section in this SI for details on the construction of the gender gaps and the sources of country-level measures of development or equality * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$