Studies in Educational Evaluation 88 (2026) 101568

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Trends in mathematics motivation between 1980 and 2015: Exploring the educational-gender-equality paradox over time

Erika Majoros

Umeå University (Department of Applied Educational Science) Naturvetarhuset, huskropp NB, Johan Bures väg 14, Umeå universitet, Umeå SE-901 87, Sweden

ARTICLE INFO

Keywords:
Educaitonal gender equality paradox
Mathematics motivation
International largescale assessments
Longitudinal analysis
Item response theory

ABSTRACT

A growing body of recent research suggests that in wealthier and more gender-egalitarian countries, gender differences in attitudes to, graduation rates in, and occupational preferences for disciplines related to science, technology, engineering, and mathematics are larger. These findings contradict the gender stratification hypothesis, i.e., that enhancing gender equality reduces the gender gaps disadvantaging girls. This study examines the educational-gender-equality paradox concerning motivation for learning mathematics using country-level panel data analysis. The analysis expands the Trends in International Mathematics and Science Study's twenty-year cognitive trend scale by including the Second International Mathematics Study. We apply a market-basket approach with item response theory modeling to construct comparable motivation scales of the student questionnaires that account for both cultural and longitudinal differences. Our results suggest that the economic development of a country is associated with the country-level gender gap in intrinsic motivation, potentially influencing pupils with a lower parental education background more.

1. Introduction

Results from the first twenty years of the Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2016) show, on the one hand, that the gender gap in mathematics achievement is closing in most countries. On the other hand, an inverse relationship was found between students' attitudes toward mathematics and the economic development of their country. This relationship is more pronounced for girls (Charles et al., 2014). Relatedly, a growing body of research suggests that in wealthier and more gender-egalitarian countries, gender differences in attitudes to, graduation rates in, and occupational preferences for disciplines related to science, technology, engineering, and mathematics (STEM) are larger (see e.g., Breda et al., 2020; Herlitz et al., 2025; Stoet & Geary, 2018). These counterintuitive differences are often referred to as *gender-equality paradoxes*.

Much of the previous research on the *educational-gender-equality paradox* (Stoet & Geary, 2018) is based on cross-sectional data from international large-scale assessments (ILSAs). However, it is possible to explore the educational-gender-equality paradox (hereinafter referred to as EGEP) employing a longitudinal design. ILSAs apply country-level trend measurement, i.e., their data are longitudinal at the system level. This allows for applying advanced statistical methods for causal inferences based on ILSA (see e.g., Gustafsson, 2008; Gustafsson & Nilsen, 2022; Robinson, 2013; Schlotter et al., 2014). Previous research has shown that it is possible to link cognitive and affective outcomes from assessments administered by the International Association for the Evaluation of Educational Achievement (IEA) before 1995 (Afrassa, 2005; Majoros, 2023; Majoros et al., 2021, 2022; Sosa Paredes & Andersson, 2025; Strietholt et al., 2013; Strietholt & Rosén, 2016) to facilitate studies exploring long-term trends (see e.g., Rosén & Gustafsson, 2016; Steinmann et al., 2023).

Another shortcoming in previous research on EGEP is the lack of problematizing the cross-cultural comparability of the attitude indicators and constructs. Moreover, the indicators used regarding EGEP vary among studies and relate to different attitudinal constructs, such as academic self-concept (e.g., Marsh et al., 2021) or intrinsic motivation (e.g., Charles et al., 2014).

The present study is set out to explore the EGEP in the long-term trends of motivation for learning mathematics. The analysis extends the TIMSS trend scale with data from 1980. We investigate whether the changes in country-level gender gaps in intrinsic and extrinsic motivation are associated with a country's development, as measured by the Human Development Index (HDI), using panel modeling. We employ a market-basket approach to achieve cross-culturally comparable attitude measures.

E-mail address: erika.majoros@umu.se.

https://doi.org/10.1016/j.stueduc.2026.101568

Available online 30 January 2026

1.1. Types of gender equality paradoxes

Depending on the field of inquiry, authors refer to several different strains of gender equality paradoxes, e.g., in relation to education, personality traits, or occupation. In the field of education, the EGEP refers to the finding that countries with high levels of gender equality have some of the widest gender gaps in secondary and tertiary STEM education (Stoet & Geary, 2018). In more gender-equal countries, the gender gap in science achievement is higher and more favorable to boys than in less egalitarian countries, as is the gap in reading, favoring girls. Boys express higher self-efficacy, more enjoyment, and a broader interest in science than do girls universally, but with a wider gap in more egalitarian countries. Yet, countries with lower levels of gender equality have relatively more women among STEM graduates.

In psychology, the *gender-equality-personality paradox* (GEPP) refers to the finding that gender differences in personality are largest in the wealthiest and most gender-equal countries (e.g., Fors Connolly et al., 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008). Studies on GEPP have operationalized the construct of personality in a number of ways, such as personal traits, values, interests, and preferences, with similar results. Fors Connolly et al. (2020) employ a longitudinal approach using data from all eight rounds of the European Social Survey across 32 European countries. They demonstrate evidence of a positive correlation between countries' level of gender equality and gender differences in values, while also finding that, for most countries, value priorities converge over time. They suggest that, instead of a causal interpretation, confounding factors potentially play a role in this relationship.

In sociology, Breda et al. (2020) refer to the gender-equality paradox related to occupation. The gender segregation across occupations is more pronounced in more egalitarian and more affluent countries. In these countries, women are highly underrepresented in math-intensive fields, even though they outnumber men in higher education. This underrepresentation contributes to gender inequality in the labor market, and it represents a loss of potential talent in a labor market with growing demand for skills related to the development of information technology and artificial intelligence.

1.2. Gender stratification hypothesis

The above paradoxes contradict the *gender stratification hypothesis* (Baker & Jones, 1993). This theory assumes that, in patriarchal societies, boys fundamentally link their academic achievements, attitudes, and ambitions to potential life opportunities. In contrast, the limited possibilities offered to girls in such societies might make them view STEM fields as less relevant to their future. The hypothesis proposes that enhancing gender equality reduces the gender gaps disadvantaging girls.

Previous research (e.g., Else-Quest et al., 2010; Fryer Jr. & Levitt, 2010; Stoet et al., 2016) is not always consistent across different datasets and gender equality indices in terms of supporting the gender stratification hypothesis (Guo et al., 2024). For instance, Else-Quest et al. (2010) show supporting evidence by demonstrating that gender equity in school enrollment, women's share of research jobs, and women's parliamentary representation are the strongest predictors of cross-national variability in gender gaps in math. They employ meta-analysis to estimate the magnitude of gender differences in mathematics achievement, attitudes, and affect using the 2003 administrations of TIMSS and Program for International Student Assessment (PISA). Stoet et al. (2016), analyzing PISA 2003 and 2012 data, show that the overall levels of mathematics anxiety of both sexes are lower in more gender equal and more developed countries, but the mean level of mathematics anxiety decreases more for boys than for girls with development. Contrary to the stratification hypothesis' prediction, they demonstrate that the proportion of mothers working in STEM was unrelated to sex differences in mathematics anxiety or performance.

Guo et al. (2024) point out that this hypothesis aligns with the central assumptions of social role theory, expectancy-value theory, and cognitive social learning theory. Social role theory presumes that the pressure on distinctive social roles is lowest in countries with more egalitarian gender roles, gender socialization, and socio-political gender equity, thereby decreasing gender differences in psychological traits, as measured with the Big Five Inventory and the Five Factor Model (Mac Giolla & Kajonius, 2019; Schmitt et al., 2008, 2017). The situated expectancy-value model (Eccles & Wigfield, 2020) assumes that individuals' choices are driven by the value they attach to available prospects and their expected accomplishment. Societal messaging may shape girls' self-perception of their interest and capabilities in STEM subjects, and in turn, dissuade women from STEM fields. Cognitive social learning theory (Bandura, 1986) posits that individuals shape their beliefs in their capabilities (i.e., self-efficacy) through observational learning and internalized societal norms. As a result, individuals align their choices more with prevailing societal gender expectations, hence perpetuating the gender inequalities.

1.3. Explanations of the gender equality paradoxes

Some of the explanations of the gender equality paradoxes stem from socio-economic reasoning. Stoet & Geary (2018, 2020) hypothesize that the relatively large gender differences in occupational interests become more clearly expressed in countries where occupational choices are less constrained by the financial incentives to study a STEM subject. Stoet & Geary (2018) analyze data from the 2015 cycle of PISA. The attitudinal constructs involved are science self-efficacy, broad interest in science, and enjoyment of science. They apply the Global Gender Gap Index (GGGI) as a measure of gender equality and employ a measure for the propensity of women to earn STEM degrees using the UNESCO graduation data.[2] They calculate personal academic strengths contrasting intraindividual mathematics, science, and reading achievement. Their findings demonstrate that the magnitude of the gender differences in personal academic strengths was strongly related to national gender equality, with larger differences in more gender-equal nations. In such countries, girls, even if they outperform boys in science or math, often achieve even higher in reading. The authors also show that countries with lower levels of gender equality had relatively more women among STEM graduates than did other countries, and that boys tend to overestimate their competence in science, even more so in countries with higher GGGI scores.

Charles et al. (2014) analyze the eighth-grade population TIMSS data from 2003 to 2011. They employ models with single indicators of intrinsic motivation for learning mathematics. The authors use the Human Development Index (HDI; Roser, 2014) as the indicator of the economic development of countries. Their findings show that the gender gap in attitudes follows a similar pattern to the gap in representation in STEM fields. In addition, their findings show more positive attitudes toward mathematics among students from less privileged social backgrounds, and that lower parental education is associated with more positive affinity and occupational aspirations only among girls. The authors suggest that lower economic status promotes more instrumental value systems on the family level.

The *resource hypothesis* argues that with greater economic development, people have more individual-level economic opportunities to make choices in line with their gendered occupational preferences (Falk & Hermle, 2018). Uunk (2023) studied whether household wealth is associated with a wider gender gap in students' math intentions using data from all countries in PISA 2012. His findings suggest that household wealth does not influence the gender gap in math intentions and show no difference between more affluent and less affluent countries. The math intentions measure is based on an item that presents a choice between being willing to study harder in math classes or language classes.

[2] http://data.uis.unesco.org

This is a conceptually different measure than those mentioned earlier from the literature or used in the present study, as it forces the respondents to choose between two subjects instead of measuring their attitude toward them independently.

Other explanations are related to gender stereotyping. Breda et al. (2020) found that gender stereotypes are stronger in more egalitarian and more developed countries. The authors compose a measure for the internalization of the stereotype that "math is not for girls" at the country level by using the 2012 PISA cycle. They use various measures for national development and equality. The authors assume that systematic gender differences in student beliefs are the product of social norms related to learning mathematics.

1.4. Critiques of the existence of gender equality paradoxes

As the above overview indicates, there is a variety of measures involved in the research on gender equality paradoxes. Some of the critiques of the existence of such paradoxes point to inappropriate measures (e.g., Richardson et al., 2020) or the lack of control measures (e.g., Marsh et al., 2021).

It is also relevant to point out the lack of problematizing the cross-cultural comparability of the attitudinal indicators and constructs. An important requirement for comparing scales in the context of ILSAs is that the latent construct is understood and measured equivalently across all educational systems (Rutkowski & Svetina, 2014). This equivalence of measurement is typically referred to as measurement invariance (Meredith, 1993), absence of measurement bias (Lord, 1980; van de Vijver et al., 2018), or lack of differential item functioning (Hambleton & Rogers, 1989).

The inconsistent support for both the gender stratification hypothesis and the gender equality paradox in previous literature warrants further investigation. We chose to focus on motivation for learning mathematics as it functions as an early factor for later gender differences in higher education choices and occupational preferences, which differs from more static traits like personality.

The present study contributes to the discussion on the EGEP by employing a country-level longitudinal design and motivation scales that take measurement invariance into account. We also address the previously mentioned lack of control measures by controlling for the achievement gap. One of the main arguments for the educational gender paradox is that the achievement gaps are closing, yet the motivation gaps do not follow this trend. In our analysis, controlling for the achievement gap warrants that we separate the effect of HDI on the motivation gaps from the achievement differences. In addition, some previous studies suggest that the EGEP varies among students with different family backgrounds. Therefore, this study involves modeling the relationship between the countries' development and motivation gaps for test-takers separately for lower and higher parental education backgrounds.

The study aims to address the following research questions:

RQ1: Does HDI significantly predict the gender gap in intrinsic or extrinsic motivation for mathematics?

RQ2: Does the achievement gap control alter the above results?

RQ3: Does the relationship between HDI and the gender gap in motivation for mathematics present the same way for test-takers with different parental education backgrounds?

2. Method

2.1. Data

The analyses were done using data from three time points: (1) the Second International Mathematics Study (SIMS 1980) administered between 1976 and 1982, the TIMSS survey administered (2) in 1995, and (3) in 2015. All surveys were administered by the International Association for the Evaluation of Educational Achievement (IEA). We used student achievement and questionnaire data of the populations representing 13-year-olds (SIMS 1980) and eighth-grade students (TIMSS cycles). The SIMS 1980 data were downloaded from the Center for Comparative Analyses of Educational Achievement website[3] (COMPEAT). Data and documentation for TIMSS 1995 and 2015 were downloaded from the IEA Study Data Repository.[4] Data management was done with the IEA IDB Analyzer (IEA, 2025), SPSS 25 (IBM Corp., 2017), and the programming language R (R Core Team, 2025).

Twenty educational systems participated in the assessment in 1980, 42 in 1995, and 40 in 2015. To maximize information, we performed two selection processes for the included countries in this study. Firstly, in Panel A, we selected data from the 11 educational systems that sampled the same grade in 1980 and 1995: Belgium (French), England, France, Hong Kong (grade 7), Hungary, Israel, the Netherlands, New Zealand, Sweden (grade 7), Thailand, and the United States. This selection ensured that we use the highest number of participating countries common with SIMS across TIMSS administrations. Since the unit of analysis in the panel regression is the educational system, sample size was an important factor in the selection process. Secondly, in Panel B, we chose the 18 systems that participated in 1995 and 2015: Australia, Canada, England, Hong Kong (grade 8), Hungary, Iran, Ireland, Israel, Italy, the Republic of Korea, Lithuania, New Zealand, the Russian Federation, Singapore, Slovenia, Sweden (grade 8), Thailand, and the United States. This selection was made for two reasons. First, to cover a similar time period to Panel A, and second, to ensure the largest sample size. The sample sizes are shown in Table A1 in Appendix A.

2.2. Variables

2.2.1. Motivation gaps

The main dependent variables are the gender gaps in motivation for learning mathematics based on the student questionnaires. Following the model proposed by Eccles & Wigfield (2002), we differentiate two domains of motivation by their source: intrinsic and extrinsic motivation. Intrinsic motivation means to engage in an activity for enjoying it, while extrinsic motives are related to instrumental reasons, e.g., future occupational goals, or meeting parental expectations.

We employ a market-basket approach described by Zwitser et al. (2017) and demonstrated in e.g., Majoros et al. (2022) for constructing the longitudinal motivation scale. Majoros et al. (2022) compare three methods for trend scaling intrinsic and extrinsic mathematics motivation scales in SIMS and TIMSS from 1980 to 2015, among which one is the market-basket approach. This approach expresses endorsement as an observed or expected score on a market basket (Mislevy, 1998) of representative items, i.e., the overall item pool over survey administrations. The main assumption is that, for each administration, the set of items included in the survey accurately represents the construct. To deal with missing data, which mainly occurs because of changes to the questionnaires, we fit an item response theory (IRT) measurement model to generate plausible responses and report plausible sum scores. To account for potential cross-cultural bias, a separate model per country is fit per panel. The IRT modeling was performed with the R package mirt (Chalmers, 2012) for the programming language R (R Core Team, 2025).

The items included in the students' questionnaires are shown in Tables B1 and B2 in Appendix B. Some questions are similar over time, but until 2011, there was no intention for trend scaling of the non-cognitive outcomes in TIMSS. Several items have been rephrased, and the number of items has been changed in the scales. The overview in Majoros et al. (2022), Table 4, shows how the surveys varied in terms of the number of items over time. The time points selected for this study,

[3] https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/studies-before-1995

[4] https://www.iea.nl/data-tools/repository/timss

beyond the sample size considerations mentioned earlier, were also those including the most indicators per construct. Table 1 shows the number of unique and common items included in this study.

One example of similar indicators is the item asking students about parental expectations. In 1980, the wording was "My parents really want me to do well in mathematics." In 1995, it was phrased "I need to do well in mathematics to please my parents," and in 2015, "My parents think that it is important that I do well in mathematics". Taking these changes into account, we apply the market-basket approach with the assumption that the items included in the survey at each time point represent the construct.

Students were presented with four response options in all items throughout all TIMSS cycles: strongly agree, agree, disagree, and strongly disagree (the wording refers to 1995). However, in SIMS, they also had a middle option: undecided. It is beyond the scope of this study to evaluate the effect of having a middle response option, but essentially, it means no information on the item. We recode the middle answers randomly and check the factor scores estimated with a single-group measurement model for the complete SIMS data. Pearson's coefficients on the individual level show strong correlations between the scales including the middle option and the recoded responses: $r = 0.999$ ($p < 0.001$) for intrinsic motivation and $r = 0.998$ ($p < 0.001$) for extrinsic motivation.

The motivational scales are established in three main steps. Firstly, we fit one measurement model per country per construct, using concurrent calibration with the pooled data of the two timepoints in Panel A. This step links the two administrations in Panel A while accounting for potential cultural differences. Model fit is evaluated with item fit plots. In the measurement model, the item parameters are estimated using the generalized partial credit model (GPCM; Muraki, 1992). The fundamental equation (Equation 1) of this model gives the probability that a student with proficiency θ_s will have, for item i, a response x_{is} that is scored in the l^{th} of m_i ordered score categories as:

(Equation 1)

$$P_{is}\left(x_{is} = l|\theta_s, \quad b_i, a_i, \quad d_{i,l}, \ldots, d_{i,m_i-1}\right) = \frac{\exp[\sum_{y=0}^{1} a_i(\theta_s - b_i + d_{i,v})]}{\sum_{g=0}^{m_i-1} \exp[\sum_{y=0}^{g} a_i(\theta_s - b_i + d_{i,v})]}$$

in which

- x_{is} is the response of student s to item i (0 or 1 if correct),
- θ_s is the ability of student s,
- a_i is the slope/discrimination parameter of item i,
- b_i is the location/difficulty parameter of item i,
- m_i is the number of response categories for item i, and
- $d_{i,l}$ is the category l threshold parameter of item i.

Secondly, after finding no indication of model-data misfit, using the item parameters, missing responses are imputed five times, thereby estimating five plausible sum scores per person per construct per panel. Thirdly, the plausible scores are transformed onto a common scale in three steps: (1) We transform the scale of SIMS scores to a mean of 10 and a standard deviation of 2 for each set of plausible scores. (2) We transform the 1995 scale of the five plausible scores with the same transformation constants. (3) We put the Panel B scores on the same scale by first matching the score distribution of the scores common year-country-grade units (grade 8 in Hungary, Israel, Thailand, and the United States) and transforming all scores in Panel B (per year) with the same constants. These steps ensure that the scores are on the same, more easily interpretable scale while keeping separate score distributions per year. Moreover, the comparisons between countries are based on a sum score over the market basket of representative items and not on estimated latent variables. In this way, the comparability across countries is not threatened by differences between them.

Table 1
Number of selected items in the affective scales of the respective studies.

	Intrinsic Motivation Scale		Extrinsic Motivation Scale	
	Unique items	Common items	Unique items	Common items
Panel A	11	2	9	3
Panel B	7	4	5	5

The country-level motivation means are achieved by aggregating the five plausible mathematics motivation scores per test-taker per scale per year to the system level, following Rubin's (1987) rules for pooling across datasets, achieved by multiple imputation outlined by Gonzalez (2014). This procedure ensures that both sampling noise and the uncertainty due to missing data are taken into account in calculating the standard errors of the means. The means are weighed with senate weights that sum to 500 for each country's data. Then the standardized difference between the means of boys and girls represents the gender gap in a country. The system-level gender gaps are calculated by extracting the boys' mean from the girls' mean.

2.2.2. Human development index

We employ the HDI, a composite statistic of life expectancy, education level, and per capita income indicators (Herre & Arriagada, 2023), as the independent variable for each country in the models at a given time point, i.e., 1980, 1995, and 2015. HDI is published by the United Nations Development Programme (UNDP).[5] The index is composed of four indicators: life expectancy at birth, expected years of schooling (for children of school entering age), average years of schooling (for adults aged 25 and older), and Gross National Income (GNI) per capita. The HDI score is then calculated by normalizing and aggregating these indicators, placing each country on a scale from 0 to 1.

The first global measures related to gender equality were the Gender Development Index (GDI) and the Gender Empowerment Measure (GEM), which were introduced in 1995 in the Human Development Report (UNDP, 1995). The GDI measures development in the same dimensions as the HDI, refining it by accounting for gender inequalities in its components, while the GEM is meant to measure the gender distribution of power in the political and economic spheres. The World Economic Forum's GGGI calculation involves converting data into female-to-male ratios for four key areas: economic participation and opportunity, educational attainment, health and survival, and political empowerment. For a comprehensive overview and evaluation of global gender equality indices, see, e.g., Hawken and Munck (2013) and Hsu and Kovacevic (2015). In this study, we employ historical IEA data from 1980 that precedes the introduction of the above-mentioned indices. Dilli et al. (2019) created a composite index measuring gender equality from 1950 to 2003, of women's to men's ratios of life expectancy, sex ratios at ages 0–5, marriage ages, parliamentary seats, average years of schooling, and labor force participation. Hence, this measure is not directly comparable to the previously mentioned indices; therefore, we chose to use HDI as a measure of countries' development.

2.2.3. Achievement gaps

To calculate the achievement gaps, we use the SIMS achievement scale scores previously linked to the TIMSS reporting scale (Majoros, 2023) and the mathematics achievement scores in the TIMSS datasets. Hence, the five plausible scores per test-taker in the selected samples are aggregated to the system level. The country means by gender are computed following Rubin's (1987) rules. The means are weighted with senate weights that sum to 500 for each country's data. Then the standardized difference between the mean achievement of boys and girls represents the gender gap in a country, calculated by extracting the

[5] https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

boys' mean from the girls' mean.

2.2.4. Parental education

To account for the students' family background, we used student questionnaire data on their mother's and father's highest educational level. Other common socioeconomic status indicators, such as home possessions, are not available at every time point. While students' reports on their parents' education might raise some concern, older children tend to report more accurately. The agreement between parents and children is best for parental occupation and least good for books in the home, while mixed for parental education (Jerrim & Micklewright, 2014).

The parental education scales in the questionnaires differ across administrations: in SIMS, the response options are (1) little or none, (2), primary school, (3) secondary school, (4) beyond secondary; in TIMSS 1995 (1) primary school, (2) some secondary school, (3) secondary school, (4) some vocational/ technical education after secondary school, (5) some university, (6) university; in TIMSS 2015 (1) some primary or lower secondary or did not go to school, (2) lower secondary, (3) upper secondary, (4) post-secondary, non-tertiary, (5) short-cycle tertiary, (6) bachelor's or equivalent, (7) postgraduate degree.

Therefore, we create a composite measure from two items inquiring about the father's and mother's educational level for each country per year via factor analysis with polychoric correlations by estimating individual factor scores. The factor scores are estimated with the R package psych (Revelle, 2025) for the programming language R (R Core Team, 2025). Based on the factor scores, we create two groups for each country-year unit: (1) below and equal to the median value, and (2) above the median value. This decision is based on insufficient variance for creating more than two groups in some units in the sample.

2.2.5. Missing data

The data preparation involved excluding some cases, i.e., students who only used middle responses on the questionnaires in SIMS. This step involved excluding 4.41 % of the original sample. The two items inquiring about the father's and the mother's education contain a considerable amount of missing data, shown in detail in Appendix C, Table C1. The data from England contained all missing values on both parental education variables; therefore, we excluded this country from the analysis concerning the subgroups.

2.3. Empirical models

We explore the relationship between motivation gaps with HDI by employing panel data analysis. Panel data techniques address the issue of unobserved heterogeneity, aiming at controlling for unobserved or omitted variables that possibly bias estimation (Croissant & Millo, 2019). The panel datasets contain the system-level gender gaps of extrinsic and intrinsic motivation, mathematics achievement, and the HDI value of the countries in the corresponding year. Panel A contains data from 11 countries in the years 1980 and 1995, and Panel B involves data from 18 countries in the years 1995 and 2015. We apply linear panel models with two-way fixed effects to control for time-invariant factors specific to each country-year unit and the time-specific effects that are common to all individual units, providing a stronger causal link. The analysis was done with the R package plm (Croissant & Millo, 2008) for the programming language R (R Core Team, 2025).

The modeling strategy involves four main models. In Model 1, we test the effect of HDI on the intrinsic motivation (IM) gap, and in Model 2, we introduce the achievement gap as a control variable. The following formula (Equation 2) describes Model 1

(Equation 2)

$$IMgap_{ct} = \quad \alpha HDI_{ct} + u_c + u_t + v_{ct}$$

in which

- *IMgap* is the country-level gender gap in intrinsic motivation
- for country c
- at time t,
- u_c is the country fixed effect,
- $u_t$is the time fixed effect, and
- v_{ct} is the error term.

Equation 3 describes Model 2 by adding the control variable $COGgap_{ct}$ for country c at time t.

(Equation 3)

$$IMgap_{ct} = \quad \alpha HDI_{ct} + \beta COGgap_{ct} + u_c + u_t + v_{ct}$$

In Model 3, we estimate the effect of HDI on the extrinsic motivation (EM) gap, introducing the achievement gap as a control in Model 4. All of these four models are fit on Panels 1 and 2. For additional sensitivity analysis, we fit four more models (Models 5–8), in which we use the log of GDP[6] per capita instead of HDI. Finally, we run all of the above models on the two groups (*lower and higher paredu*) created based on the parental education score.

3. Results

3.1. Descriptive statistics

3.1.1. Intrinsic motivation gaps

Across countries, girls are less motivated to learn mathematics for intrinsic reasons than boys, and the gap is widening over time. Table 2 presents summary statistics of the IM gaps. Negative values mean that boys show higher average motivation than girls. In almost all countries, girls' intrinsic motivation for mathematics is lower compared to that of boys on average.

Fig. 1 shows the intrinsic motivation (IM) gaps between 1980 and 1995, while Fig. 2 shows the IM gaps between 1995 and 2015. In each figure, the first graph shows the pooled data, while the second shows those in the lower parental education group, and the third shows those in the higher parental education group. Interestingly, the gender gap was close to zero on average in 1980, and in some countries, girls showed higher intrinsic motivation than boys, especially girls with more educated parents.

3.1.2. Extrinsic motivation gaps

Concerning extrinsic motivation (EM), we can observe a different trend. The gaps on average were the smallest in 1980, similar to intrinsic motivation. Boys show higher extrinsic motivation for learning mathematics than girls on average, across all time points. The gap is larger for students with more educated parents, except for 1980, when the mean gap was almost zero, and girls with more educated parents showed higher extrinsic motivation in many countries. Until 1995, the gender gaps were growing on average, with a more pronounced increase for students with more educated parents. In this subgroup, boys show

higher motivation than girls. The data suggest a trend toward zero in more recent times, albeit with boys still showing higher motivation. Table 3 presents the summary statistics of the EM gaps.

The country that stands out most from the others is Thailand, where girls tend to be more extrinsically motivated for learning mathematics than boys, as shown in Figs. 3 and 4. These figures present the EM gap changes between 1980 and 1995 and 1995–2015, respectively. Between 1980 and 1995, in most countries, the EM gap widened in favor of boys' motivation, then from 1995 to 2015, the gaps tended towards zero.

This is a similar pattern to the achievement gap. In most countries in the SIMS sample, girls outperformed boys in mathematics achievement, regardless of their parental education background. Then the

[6] https://data.worldbank.org/indicator/ny.gdp.pcap.cd

Table 2
Summary statistics of the intrinsic motivation gender gaps.

	Pooled			Lower			Higher		
	n (countries)	Mean (SE)	σ^2	n (countries)	Mean (SE)	σ^2	n (countries)	Mean (SE)	σ^2
1980	47070 (11)	-0.015 (0.051)	0.028	29924 (11)	-0.016 (0.048)	0.025	13096 (11)	-0.002 (0.052)	0.031
1995a	36412 (11)	-0.129 (0.039)	0.016	14695 (10[a])	-0.117 (0.049)	0.024	9754 (10[a])	-0.139 (0.028)	0.008
1995b	68313 (18)	-0.058 (0.026)	0.012	29408 (17[a])	-0.019 (0.032)	0.017	21829 (17[a])	-0.100 (0.028)	0.013
2015	105700 (18)	-0.173 (0.020)	0.008	32964 (18)	-0.176 (0.026)	0.012	29483 (18)	-0.171 (0.024)	0.011

Note. 1995a refers to the subset of 11 countries common with 1980, and 1995b refers to the subset of 18 countries in common with 2015. Negative values mean an advantage for boys.
[a] Without England.

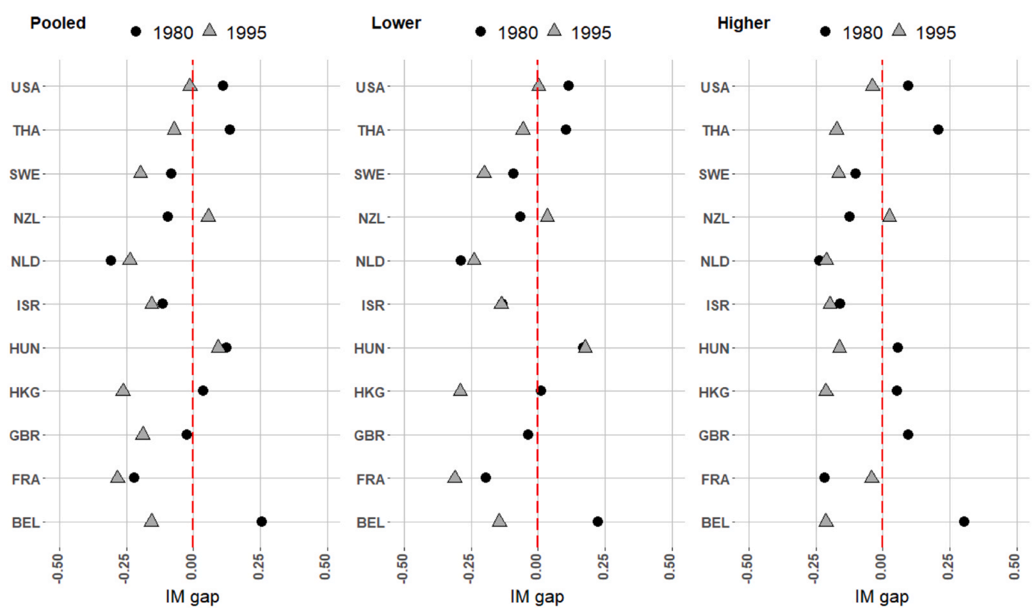

Fig. 1. Intrinsic motivation gender gap between 1980 and 1995. Note. The graph titled Pooled presents the gender gaps in the pooled data, while Lower shows those in the lower parental education group, and Higher shows those in the higher parental education group. Negative values mean that boys show higher average motivation than girls.

achievement gap became negative or close to zero in 1995 in most countries, and the narrowing continued until 2015, especially for students who have a higher parental education background, as shown in Appendix D, in Figures D1 and D2.

3.1.3. Correlations

The relationship between the variables involved in the modeling varies across panels. The HDI measure is highly correlated with (the log of) GDP per capita ($r = 0.923$ in Panel A and 0.924 in Panel B, both significant with $p < 0.001$). In Panel A, the EM gap shows a stronger correlation ($r = -0.533$; $p = 0.011$) with HDI than the IM gap ($r = -0.385$; $p = 0.078$), and the latter is not statistically significant. The achievement gap is significantly correlated with HDI ($r = -0.439$; $p = 0.041$). In Panel B, the IM gap shows a strong and significant relationship with HDI ($r = -0.610$; $p < 0.001$), while the EM gap is not significantly correlated ($r = -0.244$; $p = 0.151$) with HDI, similarly to the achievement gap ($r = 0.041$; $p = 0.712$). The Pearson's correlation coefficients are represented in Fig. 5.

3.2. Panel regression results

The panel model results shown in Table 4 suggest that HDI has a significant negative effect on the change of the intrinsic motivation gap in Panel A, i.e., from SIMS 1980 to TIMSS 1995, even after controlling for the achievement gap. Negative values of the gap indicate an advantage for boys, which means that with increasing HDI, the average level of boys' intrinsic motivation becomes higher than that of girls. As noted earlier, this decreasing trend of the IM gap is apparent in most involved countries, shown in Fig. 3. However, results from the models fit on the subgroups show that in our data, this effect only concerns students with a lower than median level of parental education background, even after controlling for the achievement gap. In Panel B, HDI does not have a significant relationship with the gender gap.

Since HDI is a composite index of life expectancy, education level, and per capita income indicators, we wanted to see how economic development alone predicts the change in these gender gaps. The panel model results presented in Table 5 suggest that in Panel A, GDP does not predict the change in the gender gap. In Panel B, GDP is a significant predictor of the change in the intrinsic motivation gender gap, and the subgroup model results suggest that it concerns students with a lower parental educational background, and the effect remains significant after controlling for the achievement gap. Growth in GDP predicts a decrease in the gender gap.

The panel model results shown in Table 6 suggest that HDI has no significant impact on the change in the extrinsic motivation gap in either panel. The same is true for GDP, as shown in Table 7.

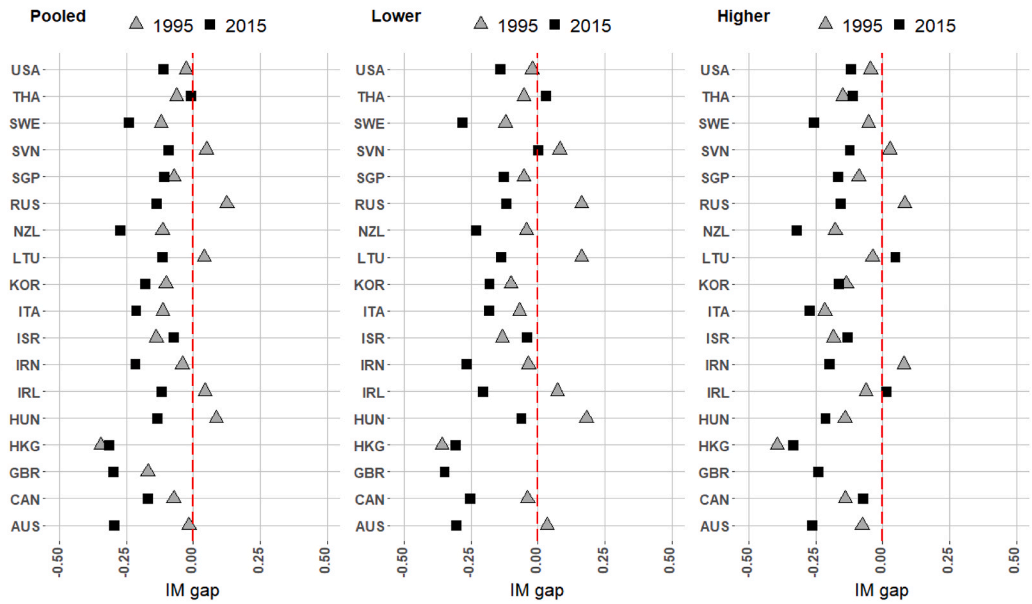

Fig. 2. Intrinsic motivation gender gap between 1995 and 2015. Note. The same logic applies as in Fig. 1.

Table 3
Summary statistics of the extrinsic motivation gender gaps.

	Pooled			Lower			Higher		
	n (countries)	Mean (SE)	Var.	N (countries)	Mean (SE)	Var.	n (countries)	Mean (SE)	Var.
1980	47070 (11)	-0.029 (0.048)	0.025	29924 (11)	-0.033 (0.044)	0.021	13096 (11)	-0.006 (0.060)	0.040
1995a	36412 (11)	-0.152 (0.052)	0.029	14695 (10[a])	-0.116 (0.067)	0.045	9754 (10[a])	-0.172 (0.042)	0.017
1995b	68313 (18)	-0.121 (0.029)	0.015	29408 (17[a])	-0.072 (0.033)	0.018	21829 (17[a])	-0.188 (0.031)	0.017
2015	105700 (18)	-0.046 (0.032)	0.018	32964 (18)	-0.043 (0.036)	0.023	29483 (18)	-0.072 (0.028)	0.014

Note. The same logic applies as in Table 2.
[a] Without England.

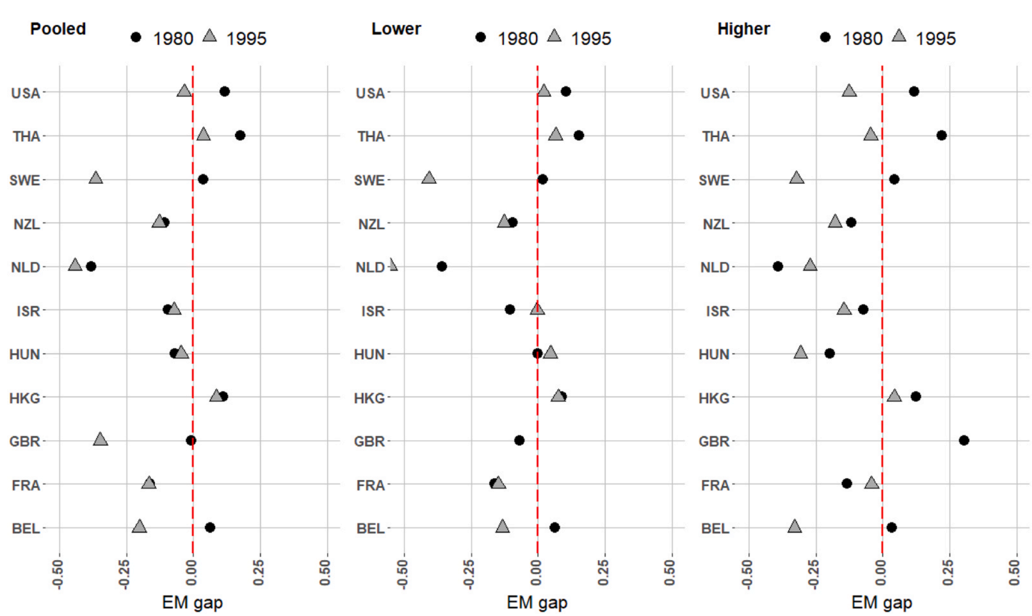

Fig. 3. Extrinsic motivation gender gap between 1980 and 1995. Note. The same logic applies as in Fig. 1.

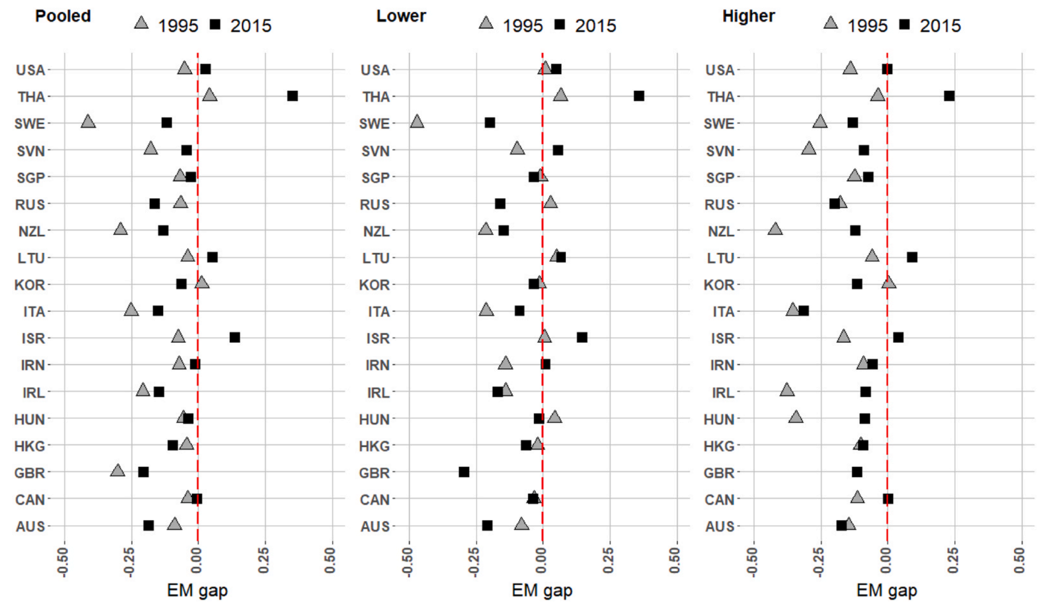

Fig. 4. Extrinsic motivation gender gap between 1995 and 2015, Note. The same logic applies as in Fig. 1.

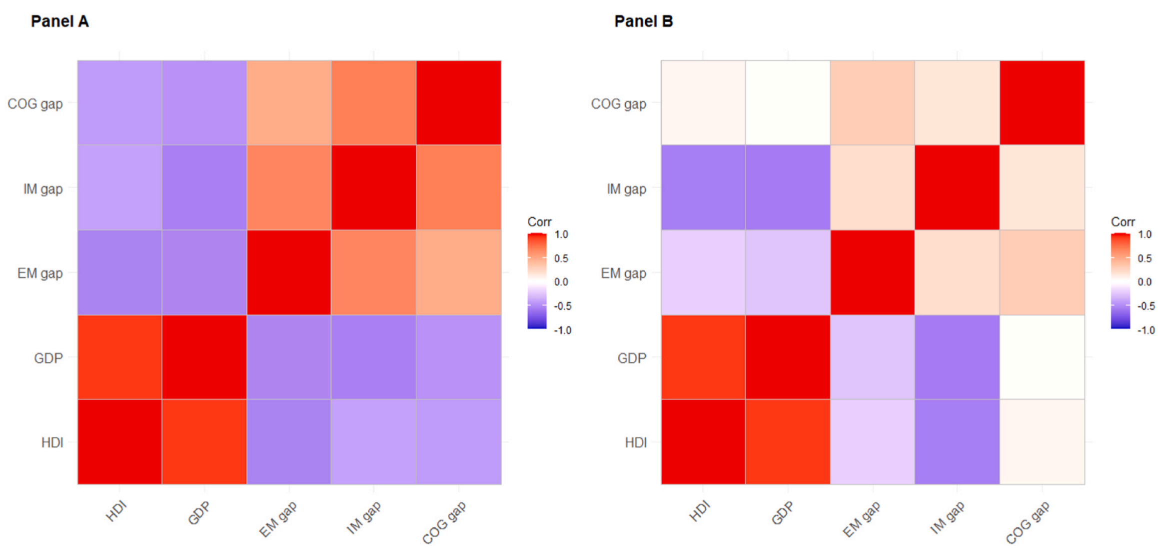

Fig. 5. Heatmaps of the Pearson's correlation coefficients in Panels A and B.

Table 4
The effect of HDI on the IM gap.

	Pooled				Lower parental education				Higher parental education			
	Estimate (SE)	*t-value*	*p-value*	R^2	*Estimate (SE)*	*t-value*	*p-value*	R^2	*Estimate (SE)*	*t-value*	*p-value*	R^2
Model 1 ST95 (n = 11)				.358	(n = 10)[a]			.450	(n = 10)[a]			.091
HDI	**-3.905 (1.439)**	-2.713	.024		**-4.057 (1.286)**	-3.155	.014		-2.719 (2.885)	-0.943	.373	
Model 2 ST95 (n = 11)				.565				.557				.339
HDI	**-4.209 (1.144)**	-3.681	.006		**-1.396 (0.441)**	-3.169	.013		-3.018 (2.227)	-1.355	.218	
Achievement gap	0.468 (0.290)	1.614	.145		0.208 (0.248)	0.837	.427		0.558 (0.371)	1.506	.176	
Model 1 T95T15 (n = 18)				.018	(n = 17)[a]			.004	(n = 17)[a]			.017
HDI	0.362 (0.567)	0.639	.532		0.240 (0.822)	0.292	.774		0.423 (0.815)	0.518	.612	
Model 2 T95T15 (n = 18)				.191				.104				.017
HDI	0.029 (0.584)	0.049	.961		-0.193 (0.906)	-0.213	.835		0.412 (0.747)	0.552	.590	
Achievement gap	0.272 (0.163)	1.670	.116		0.225 (0.156)	1.440	.172		0.009 (0.235)	0.039	.970	

Notes. ST95 refers to the joint data of SIMS 1980 and TIMSS 1995 from 11 educational systems. T95T15 refers to the joint data TIMSS 1995 and 2015 from 18 educational systems. SEs are robust standard errors (HC1) clustered by country. [a] Without England.

These results seem to partially confirm the EGEP, i.e., as a country's economy develops, the eighth-grade girls' intrinsic motivation for mathematics gets lower compared to boys on average. It is important to highlight that this effect seems to concern students with less educated

Table 5
The effect of GDP on the IM gap.

	Pooled				Lower parental education				Higher parental education			
	Estimate (SE)	t-value	p-value	R^2	Estimate (SE)	t-value	p-value	R^2	Estimate (SE)	t-value	p-value	R^2
Model 5 ST95 (n = 11)				.132	(n = 10)[a]			.122	(n = 10)[a]			.156
GDP	-0.212 (0.096)	-2.213	.054		-0.183 (0.110)	-1.654	.137		-0.308 (0.132)	-2.323	.049	
Model 6 ST95 (n = 11)				.248				.167				.298
GDP	-0.182 (0.104)	-1.755	.117		-0.163 (0.116)	-1.402	.204		-0.218 (0.163)	-1.338	.223	
Achievement gap	0.354 (0.380)	0.931	.379		0.177 (0.295)	0.601	.567		0.442 (0.381)	1.158	.285	
Model 5 T95T15 (n = 18)				.261	(n = 17)[a]			.435	(n = 17)[a]			.001
GDP	**-0.143 (0.063)**	-2.267	.038		**-0.246 (0.065)**	-3.775	.002		0.012 (0.095)	0.125	.902	
Model 6 T95T15 (n = 18)				.398				.572				.003
GDP	-0.129 (0.063)	-2.019	.061		**-0.257 (0.061)**	-4.222	<.001		0.010 (0.097)	0.098	.923	
Achievement gap	0.234 (0.123)	1.901	.077		**0.247 (0.112)**	2.190	.046		0.029 (0.239)	0.120	.906	

Note. The same logic applies as in Table 4.
[a] Without England.

Table 6
The effect of HDI on the EM gap.

	Pooled				Lower parental education				Higher parental education			
	Estimate (SE)	t-value	p-value	R^2	Estimate (SE)	t-value	p-value	R^2	Estimate (SE)	t-value	p-value	R^2
Model 3 ST95 (n = 11)				.041	(n = 10)[a]			.001	(n = 10)[a]			<.001
HDI	-1.250 (1.511)	-0.827	.430		-0.230 (1.324)	-0.174	.867		0.027 (1.743)	0.016	.988	
Model 4 ST95 (n = 11)				.094				.002				.092
HDI	-1.395 (1.389)	-1.004	.345		0.261 (1.298)	-0.201	.846		-0.112 (1.600)	-0.070	.946	
Achievement gap	0.224 (0.313)	0.715	.495		0.028 (0.275)	0.103	.921		0.260 (0.238)	1.092	.311	
Model 3 T95T15 (n = 18)				.007	(n = 17)[a]			<.001	(n = 17)[a]			.006
HDI	-0.271 (0.666)	-0.406	.690		0.022 (0.732)	-0.030	.976		-0.273 (0.703)	-0.388	.703	
Model 4 T95T15 (n = 18)				.035				.076				.006
HDI	-0.433 (0.638)	-0.679	.507		-0.392 (0.674)	-0.582	.570		-0.272 (0.752)	-0.362	.723	
Achievement gap	0.133 (0.169)	0.785	.445		0.192 (0.158)	1.216	.244		-0.001 (0.160)	-0.005	.996	

Note. The same logic applies as in Table 4.
[a] Without England.

Table 7
The effect of GDP on the EM gap.

	Pooled				Lower parental education				Higher parental education			
	Estimate (SE)	t-value	p-value	R^2	Estimate (SE)	t-value	p-value	R^2	Estimate (SE)	t-value	p-value	R^2
Model 7 ST95 (n = 11)				.107	(n = 10)[a]			.182	(n = 10)[a]			<.001
GDP	0.181 (0.141)	1.288	.230		0.228 (0.163)	1.401	.199		0.005 (0.171)	0.029	.978	
Model 8 ST95 (n = 11)				.174				.192				.101
GDP	0.203 (0.146)	1.392	.201		0.238 (0.172)	1.382	.210		0.063 (0.194)	0.324	.755	
Achievement gap	0.254 (0.303)	0.839	.426		0.087 (0.266)	0.329	.752		0.286 (0.259)	1.104	.306	
Model 7 T95T15 (n = 18)				.109	(n = 17)[a]			.191	(n = 17)[a]			<.001
GDP	-0.112 (0.082)	-1.368	.190		-0.160 (0.080)	-1.809	.091		0.008 (0.056)	0.137	.893	
Model 8 T95T15 (n = 18)				.118				.276				.001
GDP	-0.108 (0.084)	-1.278	.221		-0.168 (0.088)	-1.908	.077		0.009 (0.057)	0.159	.876	
Achievement gap	0.071 (0.158)	0.451	.658		0.190 (0.133)	1.427	.175		-0.019 (0.156)	-0.124	.903	

Note. The same logic applies as in Table 4.
[a] Without England.

parents and not those students whose parents' educational level is above the median in their respective country. Our results cannot confirm the gender paradox for extrinsic motivation.

4. Discussion and conclusions

In this study, we explored a time period of 35 years in two panels, from 1980 to 1995, and from 1995 to 2015. The analysis involved students between the ages of 13–15 years old, because this is the common target population across SIMS and the TIMSS surveys. Previous research on the EGEP also involved this age group, but it would be interesting to study this phenomenon with a developmental perspective. Numerous studies have shown that intrinsic school motivation declines across grade levels, while extrinsic school motivation is rather stable (see e.g., Lepper et al., 2005).

In both panels, all countries developed according to their HDI index and GDP per capita. Most countries show a trend for an increasing gender gap in intrinsic motivation for learning mathematics. There is a statistically significant impact of HDI on this trend from 1980 to 1995 (but not later), and GDP from 1995 to 2015 (but not earlier). The effect of HDI remains significant while controlling for the achievement gap, while this is only true for the effect of GDP in the lower parental education group. These results are partially in line with the findings by Charles et al. (2014). However, their study involved cross-sectional analysis of all participating countries spanning over TIMSS 2003, 2007, and 2011, and they used HDI as a measure of development from 1995, and one or two indicators of intrinsic motivation, depending on the administration.

Moreover, our results show that the widening of the gender gaps is more pronounced among students with a lower than median level of parental education background. The impact of development on these gaps is statistically significant in this subgroup but not in the group of students with more educated parents across both panels. This differentiation of the effect is an important contribution to the literature. Further research is necessary to investigate whether this pattern is persistent across subgroups with different family backgrounds.

The socio-economic explanations for gender-equality paradoxes (see e.g., Stoet & Geary, 2018) imply that in more economically developed countries, fewer women choose STEM-related occupations because they do not have as strong financial incentives. This could suggest that with higher economic development, girls develop lower extrinsic motivation for learning mathematics than boys. However, our results suggest that, firstly, the extrinsic motivation gap is closing in recent times. Secondly, the country development indices are not statistically associated with the extrinsic motivation gaps over time. A possible explanation is that intrinsic motivation may be more related to occupational preferences for girls in economically more affluent countries than extrinsic motivation.

There are several limitations to our study. The statistical power of the results might be improved by increased sample sizes, i.e., the number of educational systems, and a larger variety of cultures would be interesting. However, we used all available data from the studies. The subgroup analysis involved information on parental education with a considerable amount of missing data. Therefore, these results are not appropriate for generalization to the involved target populations. We employed a number of assumptions during the analyses. The affective scales have been achieved under the assumption that the questionnaire items employed at each time point sufficiently represent the construct. Since the present study is a secondary analysis of previous data collections, we did not have a chance to influence the questionnaire construction. We have used the surveys that contain the most indicators of the motivational constructs. However, all surveys use self-reported data; therefore, the results might be subject to the mono-method bias.

Future research exploring the mechanism behind EGEP is necessary. It is important to bear in mind that in this study, we explored causal *effects*, i.e., estimated causal relationships between counties' development and mathematics motivation gender gaps. A causal *mechanism* behind a causal effect sheds light on the conditions under which the relationship holds (Shadish et al., 2002). As Gustafsson and Nilsen (2022) point out, the utility and the generalizability of the effect are limited without a description of the causal mechanism.

To conclude, the present study explored the educational-gender-equality paradox of motivation for learning mathematics in grade 8 students by applying country-level panel data analysis. We applied an approach to construct comparable motivation scales with both cultural and longitudinal differences taken into account. The results suggest that the economic development of the country is related to the change in the country-level gender gap in intrinsic motivation, potentially influencing pupils with a lower parental education background more. Bearing in mind findings related to the occupational gender-equality paradox (Breda et al., 2020), it seems possible that intrinsic motivation is more related to occupational preferences in economically more developed countries than extrinsic motivation. This implies that efforts to attract more girls into STEM occupations in wealthier and more gender-egalitarian countries might need to consider promoting enjoyment in these subjects.

Pupils' motivation is influenced by their teachers' motivation, through *motivation contagion*, i.e., students generate automatic goal reasoning from information related to teacher motivation, such as behavior, emotion, and situation (Wild & Enzle, 2002). Direct predictive effect of teachers' intrinsic motivation for teaching on students' intrinsic motivation for learning has been shown (e.g., Lam et al., 2009; Radel et al., 2010; Woolfolk Hoy, 2021). Zou et al. recently explored the mechanism of this effect and found that teachers' motivating style and teacher-student relationships played a chained mediating role (Zou et al., 2024).

CRediT authorship contribution statement

Erika Majoros: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Acknowledgments

In developing the ideas presented here, we have received great support from the late Monica Rosén. We are also grateful for the input of Jan-Eric Gustafsson from the University of Gothenburg, Sweden. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A

Table A1
Sample information

Year	Country	Grade	Sample size	Mean age	SD age
1980	France	8	8300	14.11	1.04
1980	Hong Kong	7	5194	13.23	0.91
1980	Hungary	8	1743	14.22	0.52
1980	Israel	8	3337	14.03	0.39
1980	Netherlands	8	5379	14.43	0.66
1980	New Zealand	8	5045	14.02	0.45
1980	Sweden	7	3371	13.91	0.35
1980	Thailand	8	3812	14.24	0.75
1980	United States	8	6417	14.13	0.50
1980	England	9	2581	14.13	0.32
1980	Belgium (French)	8	1891	14.53	0.91
1995	Australia	8	7131	13.82	0.46
1995	Canada	8	8215	14.05	0.52
1995	France	8	2872	14.30	0.76
1995	Hong Kong	7	3347	13.21	0.77
1995	Hong Kong	8	3277	14.19	0.74
1995	Hungary	8	2774	14.27	0.52
1995	Iran	8	3530	14.60	1.03
1995	Ireland	8	3019	14.43	0.47
1995	Israel	8	1333	14.08	0.39
1995	Italy	8	2317	13.99	0.54

(*continued on next page*)

Table A1 (*continued*)

Year	Country	Grade	Sample size	Mean age	SD age
1995	Korea	8	2916	14.20	0.36
1995	Lithuania	8	2485	14.25	0.45
1995	Netherlands	8	1903	14.31	0.56
1995	New Zealand	8	3127	13.00	0.38
1995	New Zealand	9	3606	14.00	0.35
1995	Russian Federation	8	3961	14.03	0.44
1995	Singapore	8	4612	14.59	0.72
1995	Slovenia	8	2697	14.77	0.41
1995	Sweden	7	4015	13.92	0.33
1995	Sweden	8	1914	14.93	0.32
1995	Thailand	8	5806	14.29	0.65
1995	United States	8	6944	14.22	0.53
1995	England	8	1776	13.06	0.30
1995	England	9	1744	14.06	0.30
1995	Belgium (French)	8	2547	14.25	0.73
2015	Australia	8	9912	13.97	0.45
2015	Canada	8	8469	13.97	0.46
2015	Hong Kong	8	4111	14.26	0.69
2015	Hungary	8	4869	14.68	0.47
2015	Iran	8	6087	14.14	0.42
2015	Ireland	8	4644	14.43	0.42
2015	Israel	8	5416	13.95	0.37
2015	Italy	8	4444	13.80	0.49
2015	Korea	8	5307	14.40	0.33
2015	Lithuania	8	4315	14.63	0.42
2015	New Zealand	9	7911	14.06	0.34
2015	Russian Federation	8	4757	14.76	0.42
2015	Singapore	8	6077	14.36	0.47
2015	Slovenia	8	4228	13.85	0.35
2015	Sweden	8	4030	14.76	0.35
2015	Thailand	8	6393	14.40	0.50
2015	United States	8	10012	14.24	0.48
2015	England	8	4718	14.07	0.30

Appendix B

Table B1
Items in the intrinsic motivation scales

Year	ID	Question/ Statement
1980	yiwant	I really want to do well in mathematics
1980	yjobuse	I would like to work at a job that lets me use mathematics
1980	yflgood	I feel good when I solve a mathematics problem by myself
1980	yhelpo	I like to help others with mathematics problems
1980	ynomore	If I had my choice I would not learn any more mathematics
1980	ychall	I feel challenged when I am given a difficult mathematics problem
1980	ynotime	I refuse to spend a lot of my own time doing mathematics
1980	yhappy	Working with numbers makes me happy
1980	yscared	It scares me to have to take mathematics
1980	yfun	I think mathematics is fun
1980	ycalm	I usually feel calm when doing mathematics problems
1995	BSBMBORE	Do you think that mathematics is boring?
1995	BSBMENJY	Do you think that you enjoy learning mathematics?
1995	BSBMLIKE	How much do you like mathematics?
1995	BSBMSELF	I need to do well in mathematics to please myself
1995	BSBMWORK	Do you think that you would like a job that involved using mathematics?
2015	BSBM17A	I enjoy learning mathematics
2015	BSBM17B	I wish I did not have to study mathematics
2015	BSBM17C	Mathematics is boring
2015	BSBM17D	I learn many interesting things in mathematics
2015	BSBM17E	I like mathematics
2015	BSBM17F	I like any schoolwork that involves numbers
2015	BSBM17G	I like to solve mathematics problems
2015	BSBM17H	I look forward to mathematics class
2015	BSBM17I	Mathematics is one of my favorite subjects
2015	BSBM20E	I would like a job that involves using mathematics

Table B2
Items in the extrinsic motivation scales

Year	ID	Question/ Statement
1980	ypwwell	My parents really want me to do well in mathematics
1980	ynouse	Most people do not use mathematics in their job
1980	ymthjob	It is important to know mathematics in order to get a good job
1980	yuseday	Mathematics is useful in solving everyday problems
1980	ynoneed	Mathematics is not needed in everyday living
1980	ypract	Most of mathematics has practical use on the job
1980	ynotnec	A knowledge of mathematics is not necessary in most occupations
1980	ygowo	I can get along well in everyday life without using mathematics
1995	BSBMFIP2	My friends think it is important for me to do well in mathematics at school
1995	BSBMJOB	I need to do well in mathematics to get desired job
1995	BSBMLIFE	Do you think that mathematics is important to everyone's life?
1995	BSBMMIP2	My mother thinks it is important for me to do well in mathematics at school
1995	BSBMPRNT	I need to do well in mathematics to please my parents
1995	BSBMSCHL	I need to do well in mathematics to get into the school I prefer
1995	BSBMSIP2	I think it is important to do well in mathematics at school
2015	BSBM20A	I think learning mathematics will help me in my daily life
2015	BSBM20B	I need mathematics to learn other school subjects
2015	BSBM20C	I need to do well in mathematics to get into the <university> of my choice
2015	BSBM20D	I need to do well in mathematics to get the job I want
2015	BSBM20F	It is important to learn about mathematics to get ahead in the world
2015	BSBM20G	Learning mathematics will give me more job opportunities when I am an adult
2015	BSBM20H	My parents think that it is important that I do well in mathematics
2015	BSBM20I	It is important to do well in mathematics

Appendix C

Table C1
Missing data information on the parental education factor score

Year	Country	Grade	Missing (%)
1980	Belgium (French)	8	35.84
1980	England - GBR	9	9.80
1980	France	8	13.15
1980	Hong Kong-CHN	7	3.50
1980	Hungary	8	14.54
1980	Israel	8	13.46
1980	Netherlands	8	9.88
1980	New Zealand	8	9.53
1980	Sweden	7	17.04
1980	Thailand	8	2.71
1980	United States	8	5.99
1995	Australia	8	28.19
1995	Belgium (French)	8	47.12
1995	Canada	8	28.65
1995	England	8	100.00
1995	England	9	100.00
1995	France	8	51.80
1995	Hong Kong	7	22.36
1995	Hong Kong	8	20.87
1995	Hungary	8	32.29
1995	Iran	8	47.92
1995	Ireland	8	22.30
1995	Israel	8	24.88
1995	Italy	8	14.00
1995	Korea	8	11.61
1995	Lithuania	8	53.03
1995	Netherlands	8	38.17
1995	New Zealand	8	41.27
1995	New Zealand	9	37.12
1995	Russian Federation	8	20.41
1995	Singapore	8	0.60
1995	Slovenia	8	14.88
1995	Sweden	7	47.52
1995	Sweden	8	38.50
1995	Thailand	8	14.59
1995	United States	8	17.91
2015	Australia	8	58.60
2015	Canada	8	49.16
2015	Hong Kong	8	36.07

(*continued on next page*)

Table C1 (*continued*)

Year	Country	Grade	Missing (%)
2015	Hungary	8	20.16
2015	Iran	8	9.42
2015	Ireland	8	37.08
2015	Israel	8	36.00
2015	Italy	8	25.37
2015	Korea	8	37.25
2015	Lithuania	8	34.28
2015	New Zealand	9	70.30
2015	Russian Federation	8	31.93
2015	Singapore	8	38.85
2015	Slovenia	8	44.74
2015	Sweden	8	53.46
2015	Thailand	8	32.46
2015	United States	8	36.92
2015	England	8	66.74

Appendix D

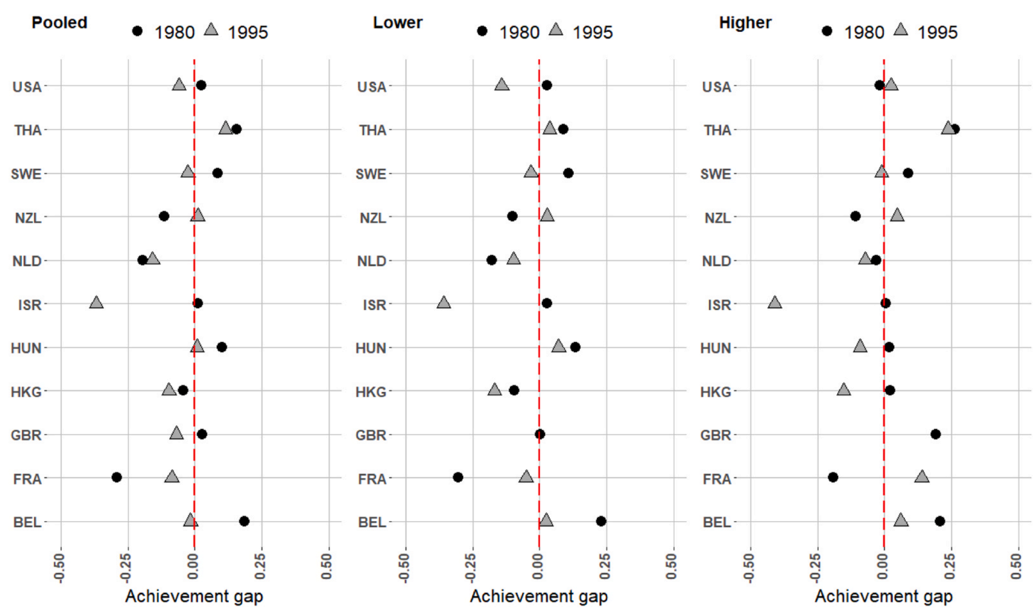

Figure D1. The achievement gender gap between 1980 and 1995

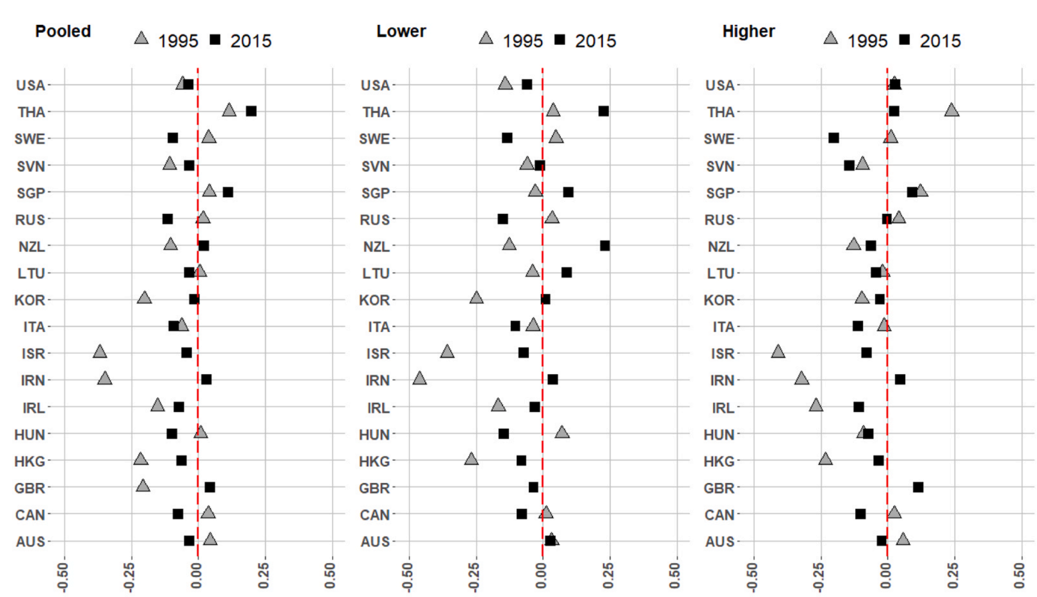

Figure D2. The achievement gender gap between 1995 and 2015

References

Afrassa, T. M. (2005). Monitoring mathematics achievement over time: A secondary analysis of FIMS, SIMS and TIMS: a Rasch analysis. In Curtis Alagumalai, D. David, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 61–77). Springer.

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education, 66*(2), 91–103. https://doi.org/10.2307/2112795

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

Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences of the United States of America, 117*(49), 31063–31069. https://doi.org/10.1073/pnas.2008704117

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6). https://doi.org/10.18637/jss.v048.i06

Charles, M., Harr, B., Cech, E., & Hendley, A. (2014). Who likes math where? Gender differences in eighth-graders' attitudes around the world. *International Studies in Sociology of Education, 24*(1), 85–112. https://doi.org/10.1080/09620214.2014.895140

Croissant, Y., & Millo, G. (2008). Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software Journal of Statistical Software, 27*(2), 1–43. https://doi.org/10.18637/jss.v027.i02

Dilli, S., Carmichael, S. G., & Rijpma, A. (2019). Introducing the historical gender equality index. *Feminist Economics, 25*(1), 31–57. https://doi.org/10.1080/13545701.2018.1442582

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61*, Article 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

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

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

Fors Connolly, F., Goossen, M., & Hjerm, M. (2020). Does gender equality cause gender differences in values? Reassessing the gender-equality-personality paradox. *Sex Roles, 83*(1–2), 101–113. https://doi.org/10.1007/s11199-019-01097-x

Fryer, R. G., Jr., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*(2), 210–240. https://doi.org/10.1257/app.2.2.210

Gonzalez, E. J. (2014). Calculating standard errors of sample statistics when using international large-scale assessment data. *Educational policy evaluation through international comparative assessments* (pp. 59–73). Waxmann Verlag.

Guo, J., Marsh, H. W., Parker, P. D., & Hu, X. (2024). Cross-cultural patterns of gender differences in STEM: Gender stratification, gender equality and gender-equality paradoxes. *Educational Psychology Review, 36*(2), 37. https://doi.org/10.1007/s10648-024-09872-3

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal, 7*(1), 1–17. https://doi.org/10.2304/eerj.2008.7.1.1

Gustafsson, J.-E., & Nilsen, T. (2022). Methods of causal analysis with ILSA data. In In. T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education*. Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_56-1.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT Area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313–334. https://doi.org/10.1207/s15324818ame0204_%25204

Hawken, A., & Munck, G. L. (2013). Cross-national indices with gender-differentiated data: What do they measure? How valid are they? *Social Indicators Research, 111*(3), 801–838. https://doi.org/10.1007/s11205-012-0035-7

Herlitz, A., Hönig, I., Hedebrant, K., & Asperholm, M. (2025). A systematic review and new analyses of the gender-equality paradox. *Perspectives on Psychological Science, 20*(3), 503–539. https://doi.org/10.1177/17456916231202685

Herre, B., & Arriagada, P. (2023). The Human Development Index and related indices: What they are and what we can learn from them. *Our World in Data*. ⟨https://ourworldindata.org/human-development-index⟩.

Hsu, Y.-C., & Kovacevic, M. (2015). *Gender equality in human development: Measurement revisited*. UNDP.

IBM Corp. (2017). *IBM SPSS Statistics for Windows*. IBM Corp.

IEA. (2025). *Help manual for the IEA IDB Analyzer (Version 5.0) [Computer software]*. IEA. ⟨www.iea.nl⟩.

Jerrim, J., & Micklewright, J. (2014). Socio-economic gradients in children's cognitive skills: are cross-country comparisons robust to who reports family background? *European Sociological Review, 30*(6), 766–781. https://doi.org/10.1093/esr/jcu072

Lam, S., Cheng, R. W., & Ma, W. Y. K. (2009). Teacher and student intrinsic motivation in project-based learning. *Instructional Science, 37*(6), 565–578. https://doi.org/10.1007/s11251-008-9070-9

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology, 97*(2), 184–196. https://doi.org/10.1037/0022-0663.97.2.184

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology, 54*(6), 705–711. https://doi.org/10.1002/ijop.12529

Majoros, E. (2023). Linking the first- and second-phase IEA studies on mathematics and science. *Large-Scale Assessments in Education, 11*(1), 14. https://doi.org/10.1186/s40536-023-00162-y

Majoros, E., Christiansen, A., & Cuellar, E. (2022). Motivation towards mathematics from 1980 to 2015: Exploring the feasibility of trend scaling. *Studies in Educational Evaluation, 74*, Article 101174. https://doi.org/10.1016/j.stueduc.2022.101174

Majoros, E., Rosén, M., Johansson, S., & Gustafsson, J.-E. (2021). Measures of long-term trends in mathematics: Linking large-scale assessments over 50 years. *Educational Assessment, Evaluation and Accountability, 33*(1), 71–103. https://doi.org/10.1007/s11092-021-09353-z

Marsh, H. W., Parker, P. D., Guo, J., Basarkod, G., Niepel, C., & van Zanden, B. (2021). Illusory gender-equality paradox, math self-concept, and frame-of-reference effects: New integrative explanations for multiple paradoxes. *Journal of Personality and Social Psychology, 121*(1), 168–183. https://doi.org/10.1037/pspp0000306

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Mislevy, R. J. (1998). Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education, 11*(1), 49–63. https://doi.org/10.1207/s15324818ame1101_%25203

Mullis, I. V. S., Martin, M. O., & Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction*. Boston College: TIMSS & PIRLS International Study Center.

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. ⟨https://www.R-project.org/⟩.

Radel, R., Sarrazin, P., Legrain, P., & Wild, T. C. (2010). Social contagion of motivation between teacher and student: Analyzing underlying processes. *Journal of Educational Psychology, 102*(3), 577–587. https://doi.org/10.1037/a0019051

Revelle, W. (2025). *psych: Procedures for psychological, psychometric, and personality research (Version 2.5.6) [Computer software]*. ⟨https://cran.r-project.org/web/packages/psych/index.html⟩.

Richardson, S. S., Reiches, M. W., Bruch, J., Boulicault, M., Noll, N. E., & Shattuck-Heidorn, H. (2020). 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. https://doi.org/10.1177/0956797619872762

Robinson, J. P. (2013). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 521–545). CRC Press.

Rosén, M., & Gustafsson, J.-E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-Scale Assessments in Education, 4*(1), 5. https://doi.org/10.1186/s40536-016-0020-8

Roser, M. (2014). *Human Development Index (HDI)*. ⟨https://ourworldindata.org/human-development-index⟩.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Schlotter, M., Schwerdt, G., & Woessmann, L. (2014). Econometric methods for causal evaluation of educational policies and practices: A non-technical guide. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 95–126). Waxmann.

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

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*(1), 168–182. https://doi.org/10.1037/0022-3514.94.1.168

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Sosa Paredes, Y. K., & Andersson, B. (2025). Linking the first, second and third IEA science studies. *Studies in Educational Evaluation, 87*, Article 101529. https://doi.org/10.1016/j.stueduc.2025.101529

Steinmann, I., Strietholt, R., & Rosén, M. (2023). International reading gaps between boys and girls, 1970–2016. *Comparative Education Review, 67*(2), 298–330. https://doi.org/10.1086/724089

Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PLOS ONE, 11*(4), Article e0153857. https://doi.org/10.1371/journal.pone.0153857

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Stoet, G., & Geary, D. C. (2020). The gender-equality paradox is part of a bigger phenomenon: Reply to Richardson and colleagues (2020). *Psychological Science, 31*(3), 342–344. https://doi.org/10.1177/0956797620904134

Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives, 14*(1), 1–26. https://doi.org/10.1080/15366367.2015.1112711

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling. *Large-Scale Assessments in Education, 1*(1), 1. https://doi.org/10.1186/2196-0739-1-1

UNDP (United Nations Development Programme). (1995). Human Development Report 1995. *UNDP (United Nations Development Programme)*.

Uunk, W. (2023). Does the gender-equality paradox hold on the micro level? An assessment of the effect of household wealth on gendered math intentions for 60 countries. *Frontiers in Education, 8*. ⟨https://www.frontiersin.org/articles/10.3389/feduc.2023.1155492⟩.

van de Vijver, F. J. R., Avvisati, F., Davidov, E., Eid, M., Fox, J.-P., Le Donné, N., Lek, K., Meuleman, B., Paccagnella, M., & van de Schoot, R. (Eds.). (2018), *Invariance analyses in large-scale studies, 201* https://doi.org/10.1787/254738dd-en

Wild, T. C., & Enzle, M. E. (2002). Social contagion of motivational orientations. *Handbook of self-determination research* (pp. 141–157). University of Rochester Press.

Woolfolk Hoy, A. (2021). Teacher motivation, quality instruction, and student outcomes: Not a simple path. *Learning and Instruction, 76*, Article 101545. https://doi.org/10.1016/j.learninstruc.2021.101545

Zou, H., Yao, J., Zhang, Y., & Huang, X. (2024). The influence of teachers' intrinsic motivation on students' intrinsic motivation: The mediating role of teachers' motivating style and teacher-student relationships. *Psychology in the Schools, 61*(1), 272–286. https://doi.org/10.1002/pits.23050

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika, 82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8