Peers, parents, and self-perceptions: The gender gap in mathematics self-assessment

Anna Adamecz[1], John Jerrim[2], Jean-Baptiste Pingault[3], and Nikki Shure[4]

[1]KRTK KTI and UCL Social Research Institute

[2]UCL Social Research Institute

[3]UCL Department of Clinical, Educational and Health Psychology and KCL Social, Genetic & Developmental Psychiatry Centre

[4]UCL Social Research Institute and IZA

Abstract

It is well established that boys perceive themselves to be better in mathematics than girls, even when their ability is the same. We examine the drivers of the gender gap in self-assessed mathematics ability using a longitudinal study of twins. Using measures of individual self-assessment in mathematics from childhood and adolescence, along with mathematics levels and test scores, cognitive skills, parent and teacher mathematics assessments, and characteristics of their families and siblings, we examine potential channels of the gender gap. Our results confirm that objective mathematics abilities only explain a small share of the gender gap in self-assessed mathematics abilities, and the gap is even larger within opposite-sex twin pairs. We find that the self-assessment of boys is positively correlated with the self-assessment of their male co-twins, not just in mathematics, but also in other abilities. However, this positive correlation is not observed between girls and their male co-twins; if anything, it is negative. This phenomenon might explain why men self-select into top jobs or STEM courses, that are filled with confident men, possibly making entry relatively easier for men. We also find that parents are more likely to overestimate boys' and underestimate girls' mathematics abilities. Gender-biased parental assessments explain a large part of the gender gap in mathematics self-assessment, highlighting the potential of the intergenerational transmission of gender stereotypes.

JEL codes: I24, J16

Keywords: gender gaps; self-assessed mathematics ability; twins; peer effects

---

We gratefully acknowledge the ongoing contribution of the participants in the Twins Early Development Study (TEDS) and their families. TEDS is supported by a program grant (MR/V012878/1) to Professor Thalia Eley from the UK Medical Research Council (previously MR/M021475/1 awarded to Professor Robert Plomin), with additional support from the US National Institutes of Health (AG046938). We are grateful to seminar and conference participants at KRTK KTI, UCL QSS, University of Loyola, ifo Institute, the Australian-wide Health & Human Capital Economics Seminar Series, and the Hungarian Society of Economics, and for Krisztina Kis-Katos and János Kiss-Hubert for their helpful comments. This study was pre-registered in the OSF Registries (https://osf.io/chv5g). This research was supported by the Economic and Social Research Council [grant number ES/T013850/1].

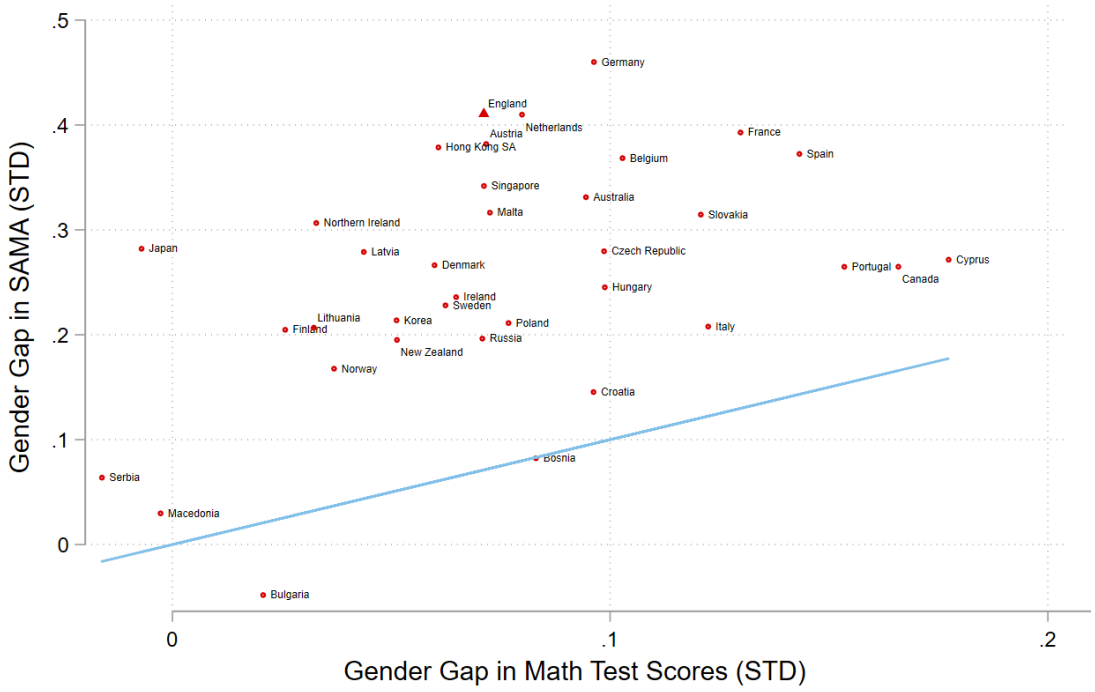
1 Introduction

Across a range of countries, contexts, and domains, men have been found to exhibit higher degrees of confidence in their ability than women (Kay and Shipman, 2014). This phenomenon has been particularly salient in the fields of science, technology, engineering, and mathematics (STEM). Not only do girls assess their mathematics ability lower than boys from an early age (Baird and Keene, 2019), but this contributes to later gender gaps in mathematics performance (Bharadwaj et al., 2016) and disparities in pay (Sterling et al., 2020). This is important since mathematics skills and participation and success in STEM fields have been linked to high labor market returns (Walker and Zhu, 2011).

Although the gender gap in mathematics performance (both grades and test scores) is narrowing in many countries, the gender gap in the self-assessment of mathematics abilities (SAMA) is still much larger. Figure 1 highlights this phenomenon using data from the most recent wave of the large-scale international assessment, Trends in International Mathematics and Science Study (Mullis et al., 2020). Almost all countries are above the 45-degree line, indicating that the gender gap in favor of boys is larger in SAMA than in mathematics performance; the magnitude of the difference in mathematics performance ranges from 0 to 0.2 standard deviations while the difference in self-assessed mathematics ability ranges from 0 to 0.45 standard deviations.

While the gender gap in mathematics performance has received much scholarly attention (e.g. Fryer and Levitt (2010)), less has been paid to the drivers of the gender gap in SAMA. Of course, the two are related, since individuals who are good at something tend to also rate their ability highly. What is perhaps worrying, however, is that the gender gap in favor of men in self-assessed ability has been shown to remain even between individuals of the same ability or when women outperform men (Ehrlinger and Dunning, 2003; Niederle and Vesterlund, 2007). This male overconfidence in their ability has been shown to explain later inequality in the labor market (Adamecz-Völgyi and Shure, 2022). Trying to understand the drivers of the gender gap in self-assessed mathematics ability is therefore important.

Figure 1: The gender gap in mathematics test scores and self-assessed mathematics ability (SAMA) internationally



Notes: SAMA and mathematics test scores have been standardized to mean zero and standard deviation one. The gender gap is calculated as the average boys' score minus the average girls' score within each country. A positive gap, therefore, denotes a gender gap in favor of boys. The 45-degree line indicates the theoretical equality of the gender gap in SAMA and in mathematics performance; in countries above the line, the gender gap in SAMA is larger than the gender gap in mathematics performance. Source: TIMSS, Grade 4 (2019)

This paper explores the drivers of the gender gap in SAMA during childhood and adolescence. We use a longitudinal study of twins from the UK that allows us to control for otherwise unobserved heterogeneity in the genetic factors, family background, and environment of boys and girls without having to worry about endogenous sex selection, birth order, or age effects. Twin data offer us a natural experiment: this is probably the only setup, where the sex of one's sibling (co-twin) is completely random. Exploiting the rich nature of the data, we estimate the gender gap in SAMA at age nine and age 12 using linear regressions conditioning on actual mathematics ability as well as a range of individual, twin-pair, and family characteristics. We draw on existing literature from education, psychology, and economics to explore the potential channels of the gender gap.

We make three contributions to the literature. First, we show that the gender gap in SAMA

persists even after controlling for mathematics grades given by teachers, mathematics test scores, measures of verbal and non-verbal cognitive abilities, birth order, birth weight, and twin fixed effects, i.e. shared genetic and environmental context. Interestingly, objective skills only explain 14-26% of the gender gap in SAMA. We document a similar gender gap in the parental assessments of children's mathematics performance, as well as in teachers' assessments, although the latter is smaller.

Second, we show that the gender gap in SAMA is even higher among opposite-sex twins than among non-related boys and girls (male and female same-sex twins). We find the gender gap in parental assessments of mathematics ability higher among opposite-sex twins, even when we control for the twins' mathematics ability. These results suggest that within opposite-sex twin pairs, there might be a stronger emphasis on who is the "mathematics person" (the boy) and the "verbal person" (the girl) within the family. This differentiation is captured in the assessments of parents and might hurt girls' confidence in their mathematics ability.

Third, we test three potential channels of the gender gap in SAMA: (1) twin peer effects; (2) parental and teachers' assessments in general, and stereotypically gender-biased parental assessments in particular; and (3) the comparative advantage of girls in English relative to math. We provide further details on these channels in the next section.

In terms of peer effects, we find that having a male co-twin (as opposed to a female-co twin) decreases SAMA, for both boys and girls alike. As the gender of one's co-twin is random, this result is causal. We do not find a significant effect of having a male non-twin sibling on average, although for girls, the magnitude of the negative relationship between having a male brother (who is not their co-twin) and SAMA is about the same as the relationship between SAMA and having a male co-twin. This highlights the importance of frame-of-reference or contrast effects for girls.

Interestingly, the mathematics performance of one's male co-twin does not contribute to the gender gap in SAMA. The SAMA of the male co-twin, however, matters, and this relationship is gender-specific. The confidence of boys is positively correlated with the confidence of their male co-twin, while between girls and their male co-twins, this positive correlation is not present.

If anything, for a girl with a male co-twin, the more confident her brother is in his mathematics abilities, the less confident she is. In other words, having a confident male co-twin seems to be good for boys but not for girls. This is true not only in mathematics but also in English (where girls perform better and exhibit higher confidence than boys) and in physical abilities (where boys are slightly more confident). These results could indicate that some of the educational and labor market gender gaps, like those in STEM studies and top jobs, might be related to this phenomenon. STEM tracks and top jobs are traditionally filled by confident men, and such a peer group might make entry easier for men and thus relatively harder for women. An important caveat of these results is that while the gender of one's co-twin is random, their self-assessment is not. Thus, while we control for various measures of individual characteristics, we cannot exclude the possibility of the over- or underestimation of these statistical relationships.

While we are not able to identify the causal effect of parental evaluations on SAMA, we find suggestive evidence that the intergenerational transmission of gender stereotypes might be important in producing the gender gap in SAMA. As mentioned above, parents also exhibit a gender bias when assessing their sons' and daughters' mathematics abilities. Even teachers exhibit a similar bias in how they assess male and female pupils. Parental assessments make a large contribution to the gender gap in SAMA: they explain 23% of the gap even when we account for the twins' actual mathematics ability. We probe this channel further by constructing a binary variable that captures whether the assessment of parents is stereotypically gender-biased, i.e. they underestimate their daughter or overestimate their son in math. We find that the largest gender gap in SAMA is among those young people with stereotypical parental assessments.

In terms of comparative advantage, we find that although those with higher performance in English have lower SAMA (hence, they are more likely to view themselves as a "verbal person"), this relationship is not gender-specific; thus, it does not contribute to the gender gap in SAMA. It is true for both genders that their (conditional) self-assessment in mathematics is positively correlated with their (conditional) self-assessment in English, and this correlation is even higher for girls. This result suggests that general confidence in abilities might be more important for girls in terms of how

they self-assess their mathematics ability.

Taken together, our results lend support for the transmission of gender biases from adults to children, and from male peers to both men and women, even though we cannot supply causal evidence in this respect. We suggest that potential interventions aiming to increase SAMA among girls and decrease the gender confidence gap, in general, should also target parents. Furthermore, as we also document a gender gap in teachers' assessments, conditional on mathematics levels that they themselves gave to their students, we suggest increasing teachers' awareness of their potentially gender-biased performance evaluations.

The rest of the paper is structured as follows. In Section 2 we elaborate on the potential channels of the gender gap in SAMA outlined in the introduction. In Section 3 we present the data used in this paper as well as some descriptive statistics. In Section 4 we outline the empirical strategy. This is followed by the results of our estimation in Section 5. Finally, in Section 6 we conclude.

2 Potential mechanisms and related literature

Drawing on the previous literature introduced above, the data allows us to test three potential channels that might affect the gender gap in SAMA. These are: (1) twin peer effects; (2) parental and teachers' assessments in general, and stereotypically gender-biased parental assessments in particular; and (3) the comparative advantage of girls in English compared to math.

As mentioned above, using data of twins supplies a natural experiment: the gender of one's co-twin is random. There is an extensive literature on peer effects, including siblings (e.g. Nicoletti and Rabe (2019)), and an individual's twin is likely to be their main point of reference or comparison (i.e. their key peer). We exploit the exogeneity of twin sex to determine if the sex of an individual's twin impacts the gender gap in SAMA. There is a literature examining the long-term effects of in-utero testosterone exposure (Auyeung et al., 2009). Bütikofer et al. (2019), for example, find that women exposed to increased testosterone in-utero via a male twin experience a lower probability of completing education and lower fertility later in life. This also holds true for women whose male

twin died shortly after birth, indicating the importance of this biological channel.

Apart from in-utero testosterone exposure, other peer effect mechanisms behind the gender gap in SAMA could include the environmental exposure to male siblings. Girls with brothers have a boy as their closest peer and the most direct point of comparison. There is a broad literature on the importance of peer effects (Sacerdote, 2011), which also highlights the importance of the gender of one's peers in the classroom (e.g. Lavy and Schlosser (2011)). Parents may also parent boys and girls differently for a variety of reasons. They may also have set gender roles within the home that reinforce societal gender stereotypes. It has been shown that growing up in families with a preference for sons decreases girls' mathematics performance (Dossi et al., 2021). We investigate these peer effects looking at the gender-specific correlation between one's own SAMA and their co-twin's mathematics ability and SAMA. These variables are interacted with gender to explore heterogeneous effects by gender.

Psychologists have pointed to the importance of gender stereotypes, where certain fields are viewed as either feminine or masculine, in determining how individuals assess their own ability in those subjects. This has its origins in social role theory, which states that gender stereotypes emerge because we observe men or women occupying certain positions in society (Eagly and Wood, 2012). There is well-documented evidence that both men and women view mathematics as a masculine subject (Makarova et al., 2019). This implies that girls may self-assess their mathematics ability lower than boys because they learn these biased assessments from the adults (e.g. teachers and parents) in their environment. When these adults are particularly gender stereotypical in how they assess children, their assessments may be even more salient. In a related paper, Nicoletti et al. (2022) show that parents assess sons' mathematics ability higher than daughters'. We explore this channel by including parental and teacher assessments in our models. We also create an indicator for whether parents assess their children's mathematics ability according to gender stereotypes and include this in the model. This variable is also interacted with gender to explore the differential effects of gender stereotypes for boys and girls.

In social psychology, people are assumed to see themselves as either a "math" person or a "ver-

bal" person, but usually not both at the same time (Marsh and Hau, 2004). Furthermore, results of consecutive rounds of the Programme in International Student Assessment (PISA), show that boys tend to somewhat outperform girls in mathematics, but girls are usually much better in reading than boys (OECD, 2020). In PISA 2018, the average gender gap in favor of girls was six times as large in reading (30 PISA points) as the gender gap in mathematics in favor of boys (5 PISA points). Theoretically, the comparative advantage of girls in English might enhance their self-assessment of being a verbal person rather than a mathematics person. This could in turn explain some of the gender gap in SAMA. Goulas et al. (2020) find that the comparative advantage of boys in STEM subjects relative to non-STEM subjects explains at least 12% of the gender gap in STEM specialization while Breda and Napp (2019) show that comparative advantage in mathematics explains 75% of the gender gap in math-intensive studies. We account for the comparative advantage of girls in English by controlling for measures of English ability as well as self-assessed English ability. These variables are also interacted with gender to explore heterogeneous impacts.

3 Data and Descriptive Statistics

We use data on twins born in the UK from the Twins Early Development Study (TEDS) (Rimfeld et al., 2019). All twin pairs born in the UK from 1994-96 (in four school cohorts) were included in the original sampling frame, and are followed from birth. Out of the four cohorts, we use data on two cohorts, born in 1994-1995. As will be detailed below, the key variables we need were collected at age nine, and the age nine data collection covered two cohorts only. The data includes rich, repeated measures of cognitive and non-cognitive skills, parental background, and educational outcomes. For our study, using TEDS offers the possibility of looking at the gender gap in self-assessed mathematics ability while controlling for shared genetic and home environments, which would not be possible in other datasets due to the endogenous nature of the decision to have multiple

children via multiple births.[1]

As mentioned above, we focus on the age nine sample because this is the first age at which SAMA was collected. It was also only at this age that parents and teachers were also asked to assess the twins' mathematics ability. The age nine data collection was restricted to twins born between January and August 1994 (Cohort 1) and twins born between September 1994 and August 1995 (Cohort 2). Our main estimation sample includes those who have non-missing data for the variables we use at age nine (3,877 individuals). This is a rather small sub-sample of the main study (15,216 individuals in Cohort 1 and 2) because we require data for both twins as well as data from their parents and teachers.

We investigate how this subsample of TEDS relates to those who either dropped out or did not provide all data that we need at age nine (11,339 individuals) in Table O1 in the Online Appendix. Furthermore, we provide robustness checks to our main results in the Online Appendix where we account for the observable selection of those in our analytic sample using three methods to create weights: probit, random forest, and entropy balancing. We model the probability that individuals are included in our analytical sample using probit and random forest models. Control variables include information collected in the first wave: parental education and measures of socioeconomic status, family structure, number of siblings, and ethnicity. We fit the individual-level estimated probabilities of being in the analytical sample from both approaches, and re-estimate our main results by using the inverse of these probabilities as estimation weights. As those included in the analytical sample differ from those who dropped out (or not reported data) (Table O1 in the Online Appendix), we apply a balancing technique, entropy balancing (Hainmueller, 2012), to construct individual-level weights to equate the moments of the distributions of these variables across the two groups. Using these entropy-balanced weights, we weigh individuals in the analytical sample in such a way that their individual characteristics have the same distribution as the individual characteristics of those who were excluded from the sample. We show in Figure O4 in the Online

[1] Twin samples are not necessarily representative of the population, which might hinder the external validity of our results. Mothers of twins tend to be on average older, higher educated and healthier, than the mothers of singletons due to IVF (Bhalotra and Clarke, 2019).

Appendix that using these weights eliminates statistical differences between those in the main sample and those who were excluded. Re-estimating our (unweighted) main results using any of these three methods leads to similar results; thus, we are confident that (observed) sample selection is not driving our results. However, we cannot exclude potential unobserved sources of sample selection.

SAMA was also collected in the age 12 sweep, which we use to provide robustness checks to our main results. We also provide a robustness check on our main model using the overlap of the age nine and age 12 samples (509 individuals).

3.1 Self-Assessed Mathematics Ability (SAMA)

TEDS measures self-assessed mathematics ability via three survey questions administered at age nine and 12. The survey asks the following three questions:

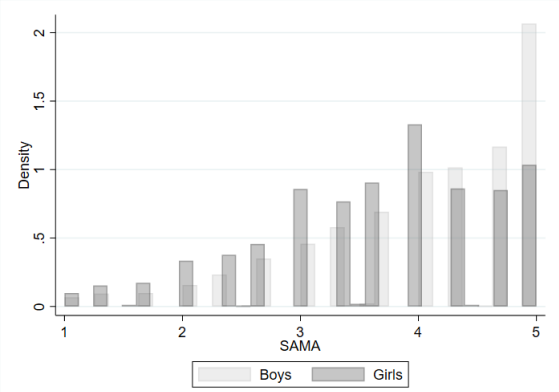
How good do you think you are at:

1. *solving number and money problems.*
2. *doing Maths in your head.*
3. *multiplying and dividing.*

There are five ordinal answers to each: very good; quite good; doing OK; not so good; not good at all, coded using a Likert scale from one (worst) to five (best). The average of responses to the three questions is provided in the data. The average SAMA at age nine is 3.83 in our analytical sample (Table A1 in Appendix A). For the purposes of our regression models, we standardize the SAMA measure to mean zero, standard deviation one so that all coefficients may be interpreted in terms of effect sizes.

Figure 2 presents the distribution of SAMA for the age nine sample by gender. Interestingly, both distributions are shifted to the right: the majority of individuals have a positive view of their mathematics abilities. This result corresponds to findings in the overconfidence literature that people are overconfident in their ability on average (Alicke et al., 2005; Dunning et al., 2004). It is

Figure 2: The distribution of self-assessed mathematics ability, age nine



Notes: N = 3,877. Source: TEDS (Rimfeld et al., 2019). The five ordinal categories are the following: 1: not good at all; 2: not so good; 3: doing OK; 4: quite good; 5: very good.

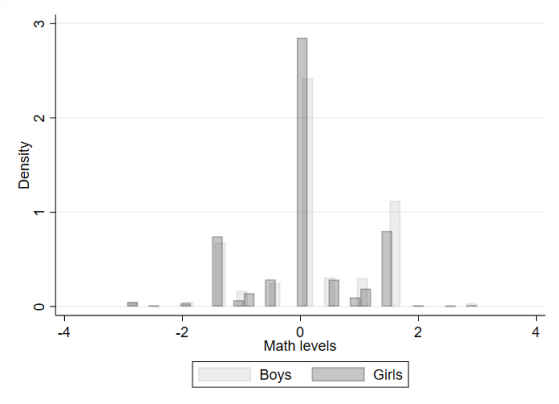
also clear that boys assess their mathematics abilities higher than girls on average. The distribution of SAMA is skewed to the right for both genders, but boys show a larger bunching at the highest self-assessment level. In our main analytical sample, the raw gender gap in SAMA at age nine is 0.382 standard deviations (Table A2 in Appendix A).

3.2 Objective Skills in Mathematics

Mathematics levels. Teachers evaluate their students' mathematics ability at ages seven, nine, and 12 according to National Curriculum levels (1 to 5) on three aspects of math: using and applying mathematics; numbers and algebra; shapes, space, and measures. This was used by the survey organizers to compute an overall sum score ranging from 3-15, which was then standardized to mean zero, standard deviation one.

Figure 3 shows the distribution of observed mathematics ability by gender. As this measure has been standardized over the total TEDS sample, the average is zero. This figure shows that boys outperform girls in mathematics at age nine. In our analytical sample, at age nine, the mathematics level of boys (0.157) is 0.129 standard deviation higher than the mathematics level of girls (0.029) (Table A2 in Appendix A).

Figure 3: The distribution of mathematics levels, age nine



Notes: N = 3,877. Source: TEDS (Rimfeld et al., 2019).

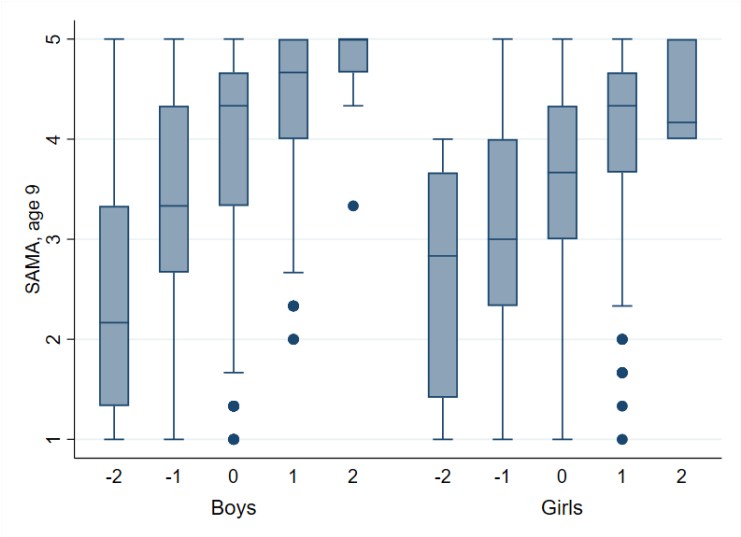
Due to being constructed from categorical variables, the distribution of mathematics levels is trimodal: about half of the distribution is around the mean, and 25-25% are below or above the mean (Figure 3). Measuring objective mathematics abilities well is key for our analysis, so we provide several robustness checks to our main results to show that measurement error does not drive our results. These robustness checks are detailed in Section 4.

Mathematics test scores. At age 12, study members also completed an Internet-based mathematics test. The scores of this test have been standardized to mean zero and standard deviation one, and follow a normal distribution (Figure A3 in Appendix A). Robustness tests using these scores are detailed in Section 4.

3.3 SAMA along the levels of mathematics abilities

Figure 4 shows the distribution of SAMA across the standardized measure of mathematics performance. At levels above average (greater than zero), the female distributions of SAMA display more variance and a lower mean than the male distributions, which indicates that even very high-achieving girls rate their mathematics ability lower than boys.

Figure 4: The distribution of self-assessed mathematics over standardized mathematics levels, age nine



Notes: N = 3,877. Source: TEDS (Rimfeld et al., 2019).

3.4 Control Variables

In addition to objective skills in math, we control for the following variables in our models:

- Gender (a dummy variable for being female): captures the gender gap.
- Cohort fixed effects: the age nine wave of TEDS covers two school cohorts, born between 1994-96. As consecutive school cohorts might differ from each other or might be exposed to different circumstances, we control for cohort fixed effects in all models.
- Cognitive abilities: TEDS measures objective cognitive abilities via tests taken at various ages. In our analysis, we use cognitive ability measures from age nine in our main models, while also providing robustness checks using cognitive ability measures from age seven and 12. To accommodate the potentially heterogeneous gender gap in different types of cognitive skills, we use two separate cognitive skill indexes, which were provided in the data: verbal skills and non-verbal skills.

- Individual characteristics that might affect mathematics outcomes and mathematics self-assessment: whether individual *i* is the elder twin (i.e., born first); whether individual *i* was heavier at birth than the co-twin; and birth weight in grams.

3.5 Potential Channels

We use the following variables to test the three potential channels outlined above.

1. Sibling peer effects

    - Having a male co-twin: as mentioned above, having a male co-twin (as opposed to having a female co-twin) could affect both girls and boys through increased in-utero testosterone exposure, as well as provide a different environment in the family.
    - Having a brother to capture the experience of growing up with male siblings (apart from one's co-twin), and potentially test whether the relationship between SAMA and having a male co-twin vs having a brother who is not a co-twin differs. Note that most siblings are older than the twins in the data,[2] which means that using whether the individual has an older brother (as opposed to just brother) would lead to similar results.
    - Twin peer effects: we look at the role of co-twin's mathematics level and SAMA, as well as their self-assessed English and physical abilities. Self-assessed English and physical abilities are captured similarly to SAMA. For English, the survey asks three questions: How good do you think you are at reading, writing, and spelling. All potential answers are coded using a Likert scale from 1 to 5, and the average of the three questions is provided in the data. For physical abilities, the survey again asks three questions: How good do you think you are at playing team games, races and competitions, and physical education classes. All potential answers are coded using a Likert scale from 1 to 5, and the average of the three questions is provided in the data.

2. Transmission of parental stereotypes

[2] The share of parents having another child after having twins is low.

- Parental (and teachers') assessments of the mathematics abilities of the twins. The questions are the same as for SAMA.
- Measure of gender-stereotypical parental assessment: we construct a binary variable that captures whether parents' assessment of their children's mathematics abilities is stereotypically gender-biased if they either:
  - Overestimate their son in math
  - Underestimate their daughter in math

  The variable is child-specific and may vary within twins/families.

  We determine over- and underestimation by comparing the mathematics levels and the parental assessments of children. First, we model the assessment category given by parents using a multinomial logit model, where we condition on objective mathematics levels as well as verbal and non-verbal cognitive skills measured at age nine. Then, we compare the category given by parents to the category predicted by the model to determine whether parents over- or underestimate their children's mathematics skills. In our main results, we use the terciles of parental assessments as the outcome variable in these models (hence we model three categories). We also provide a robustness check where instead of terciles, we use the parental assessment level on a 1-5 scale (taking the integer of the parental assessment values, that are the average levels given in response to the three questions as for SAMA) which results in a five-category model. The two methods lead to very similar results.

  The gender gap in parental assessments is presented in Table A2 in Appendix A. Boys are more likely to be overestimated while girls are more likely to be underestimated in mathematics. Overall, 26% of young people received a stereotypically gender-biased assessment from their parents (Table A1 in Appendix A).

3. Comparative advantage of girls in English compared to mathematics

- We test whether those with higher abilities in English have lower SAMA, and whether

such relationship is heterogeneous by gender. English abilities are measured similarly to mathematics abilities using National Curriculum levels from 1 to 5, given by the teachers.

- We also test whether SAMA is related to self-assessed English abilities (as a proxy for confidence in general).

Descriptive statistics of these variables are shown in Table A1 in Appendix A.

4 Empirical methods

We investigate the gender gap in SAMA using linear regression models. First, we estimate the following model:

$$SAMA_{i,j} = \alpha + \beta_{OLS} female_{i,j} + X_{i,j}\delta + u_{i,j} \tag{1}$$

Where,

j represents twin pairs

i represents the individual within a twin pair

$female_{i,j}$ captures whether individual i is female

$X_{i,j}$ is a matrix of control variables discussed in the previous section

$u_{i,j}$ is the usual error term, robust and clustered by twins.

In this model, β_{OLS}, the estimated parameter on our variable for female, captures the gender gap in the outcome variable, conditional on $X_{i,j}$.

Our preferred empirical model, however, also controls for twin-pair fixed effects (FE). Whenever possible, i.e. when we do not want to control for individual characteristics that are constant within twin pairs, we use twin-pair FE models. These models identify the gender gap within opposite-sex twin pairs and allow us to account for the shared genetic and home environment common to the twin pair. To do this, we estimate variations of the following model:

$$SAMA_{i,j} = \alpha + \beta_{FE} female_{i,j} + X_{i,j}\delta + \nu_j + u_{i,j} \tag{2}$$

Where ν_j is the twin-pair fixed effect, and all other variables are as previously outlined. β_{FE} captures the within-twin pair gender gap in the outcome variable.

We estimate our models additively, beginning with the bivariate regression of SAMA on the female dummy in Model 1. This is extended to include mathematics performance at age nine in Model 2. This allows us to examine whether boys are more confident in their mathematics ability as compared to girls who have the same level of performance. In Model 3, we introduce additional cognitive ability controls as well as individual demographic characteristics, which may drive some of the gender gap in SAMA. In Model 4, we introduce twin-pair fixed effects. This allows us to control for unobserved heterogeneity common to the twin-pair, e.g. shared genes and family environment.

We also estimate the same models on the outcome variables of teachers' and parents' assessments of the twins' mathematics skills. This allows us to probe the gender gap in the assessments made by adults of boys and girls shown in Section 3. These models follow the same logic as the aforementioned models for SAMA, but have either teacher or parent assessments as the outcome variable.

We provide the following robustness tests to our main models on SAMA. First, we re-estimate our main models treating the mathematics level variable as categorical. We do this because the mathematics level variables were constructed from three categorical variables and about 50% of observations are around the mean.

Second, we address issues of measurement error. The measurement of objective mathematics skills is key to estimating the gender gap in SAMA over and above objective mathematics performance. Furthermore, applying FE models might exacerbate any measurement error issues (Collischon and Eberl, 2020). Thus, we aim to reduce measurement error in mathematics levels in six ways. First, we also control for mathematics levels and verbal and non-verbal cognitive skills from age seven (on the overlap sample of those who participated in age seven and nine data collection). Second, as participants completed a mathematics test at age 12, we repeat the estimation on the age 12 sample (measuring SAMA at age 12) and also control for age 12 test scores on top of mathematics levels. Third, exploiting the overlap sample of the age nine and 12 data collections, we re-estimate our main model on age 12 SAMA while controlling for both age nine and age 12 mathematics levels and age 12 mathematics test scores as well. Fourth, we repeat the previous exercise by controlling for age seven, nine, and 12 level and test scores variables at the same time. Note that the overlap samples have fewer observations. Our last two methods are two instrumental variable approaches. First, following Ladd and Walsh (2002), we instrument age nine math levels by age seven math levels. Second, we follow the ORIV approach of Gillen et al. (2019), which uses both age nine math levels to instrument age seven math levels and age seven math levels to instrument age nine math levels at the same time. All these methods lead to similar results.

In our third robustness test, we investigate whether the gender gap varies along the distribution of SAMA. We treat SAMA as a categorical variable (as opposed to continuous) and estimate a multinomial logistic model.

Lastly, we re-estimate our main results on a subsample that only contains dizygotic twins. Opposite-sex twins are dizygotic by nature, so we test what happens when we exclude monozygotic twins from the analytical sample.

4.1 Exploring the channels

We explore the role of the potential channels outlined in Section 2 by extending the main model with variables accounting for the three channels as well as their interaction with the female dummy

to explore heterogeneous effects.

First, we estimate a series of models to account for sibling peer effects. These models allow us to examine the role of siblings (reference point) in the gender gap in SAMA. We do this by including a dummy variable for whether an individual's co-twin is a boy to the model without twin fixed-effects. We then introduce an interaction term for whether the individual is female and their co-twin is a boy. In a separate model, we replace having a male co-twin with having a brother (twin or not) to test whether the same relationship occurs as for having a male co-twin. Finally, we estimate this last model separately for opposite-sex and same-sex twins to investigate the consequences of having a brother separately for girls who have or do not have a male co-twin.

We further probe the peer effects explanation by including further characteristics of the twin beyond their gender: their SAMA and mathematics levels. This allows us to delve further into the reference point hypothesis and explore whether their co-twin's ability and SAMA might discourage girls and explain part of the gender gap. Lastly, we repeat this last exercise for two further facets of self-assessment: self-assessed English and physical abilities.

Second, we extend the main model with the variable capturing whether one received a stereotypically gender-biased parental assessment, as well as the interaction term of this variable with female. Again, we estimate linear models with OLS and twin-pair FE models.

Lastly, we extend the main model with objective measures of English ability. Then we add self-assessed English ability, as well as the interactions of both variables with female. We estimate these three new models using OLS and twin-pair FE models.

5 Results

5.1 Main Results

Table 1 presents the main results obtained from estimating Equation 1 on the age nine sample. In all models, the coefficient of interest is on the female dummy, indicating the difference between boys and girls. Model 1 reveals a large and statistically significant raw gender gap in SAMA of

-0.38 standard deviations. Girls rate their own mathematics ability nearly 40 percent of a standard deviation lower than boys. In Model 2, this is reduced by the inclusion of mathematics ability by five percentage points (13%), but still large (-0.33 SD) and statistically significant. This result indicates that a girl with the same mathematics skills as her male peer still rates her mathematics ability one-third of a standard deviation lower on average.

Table 1: The gender gap in mathematics self-assessment (SAMA), age nine

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 3 OS subsample	(5) Model 4
Female	-0.376***	-0.328***	-0.324***	-0.449***	-0.447***
	(0.034)	(0.032)	(0.032)	(0.051)	(0.051)
Math level, age 9		0.372***	0.327***	0.319***	0.359***
		(0.016)	(0.019)	(0.030)	(0.032)
Verbal abilities, age 9			0.054***	-0.030	0.082**
			(0.018)	(0.030)	(0.035)
Non-verbal abilities, age 9			0.061***	0.110***	0.131***
			(0.020)	(0.034)	(0.033)
Elder twin			0.038	0.062	0.034
			(0.026)	(0.051)	(0.027)
Heavier twin at birth			0.042	0.051	0.039
			(0.028)	(0.055)	(0.043)
Birth weight, grams			0.000	-0.000	0.000
			(0.000)	(0.000)	(0.000)
Constant	0.182***	0.118***	-0.064	0.251*	0.033
	(0.035)	(0.032)	(0.089)	(0.151)	(0.236)
Observations	3,877	3,877	3,877	1,186	3,877
R-squared	0.036	0.165	0.174	0.195	0.164
Twin FE	No	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	Yes	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

In Model 3, we exploit the rich nature of the TEDS data and include a range of control variables for cognitive ability as well as individual characteristics. These do very little to reduce the gender gap in SAMA (-0.32 SD; six percentage points or 16% smaller than the raw gap).

In Model 4, we restrict the sample to opposite-sex twins while in Model 5, we introduce twin-pair fixed effects on top of the aforementioned control variables. This means we estimate our gender gap within opposite-sex twin pairs as outlined in Equation 2. Interestingly, the gender gap increases in magnitude to -0.45 SD.[3]

In Table O6 in the Online Appendix, we repeat the same estimations by adding the interaction terms of female and all control variables to the model to see any potential differential effects. Returns to mathematics levels in terms of SAMA do not differ between men and women (Model 2). In the OLS model (Model 3), none of the interaction terms are statistically significant or meaningful in magnitude, while in the FE model (Model 4), the interaction term of female and verbal skills is significant and negative. Thus, within opposite-sex twin pairs, girls' SAMA is negatively correlated with their verbal abilities.

We provide the following robustness checks to support our results on the contribution of objective mathematics abilities to the gender gap in SAMA in Appendix B. First, as the distribution of mathematics levels is trimodal (Figure 3), we control for mathematics levels as a categorical variable in Column 1 and Column 4 of Table B1. This does not change the results.

Second, we try to reduce any potential measurement error in mathematics levels at age nine in various ways: controlling for mathematics levels and cognitive skills at age seven (Columns 3 and 6 of Table B1), as well as using two types of IV strategies (Tables B2 and B3). While the overlap sample of the age seven and age nine data is somewhat smaller than our main analytical sample, the conditional gender gap is similar, and not different from the earlier estimates.

We also repeat the estimation using age 12 SAMA as the dependent variable in Table B5. The age 12 raw gender gap in SAMA is similar in magnitude to the age nine gap (note that most of the age 12 sample covers different individuals as compared to the age nine sample, the overlap of the two is only 570 individuals), -0.39 standard deviation (Model 1). Controlling for age 12 mathematics levels decreases the gap by 14.5 percent to -0.34 (Model 2). Once we also control for

[3]Note that what we measure here is not girls being less confident on average than boys in general, but only in their SAMA. In self-assessed English abilities, for example, the gender gap is positive: girls assess themselves to be better than boys, even after controlling for objective abilities in English (Table O8 in the Online Appendix).

age 12 test scores and age 12 cognitive skills, the gap decreases further to -0.299 (Model 3). Thus, all age 12 mathematics and cognitive skill measures explain 24.1% of the gender gap in SAMA at age 12.

When we restrict the sample to those with both age nine and age 12 data and control for age 12 and age nine mathematics and cognitive skill measures as well, the gender gap in SAMA is still 0.34 standard deviations (Model 4). When we restrict the sample further to those with age seven, age nine, and age 12 data and control for all available measures from the three ages, the gap is still 0.29 standard deviations (Model 5). Repeating the same exercise in twin FE models yields similar results (Table B6), as well as restricting the sample to dizygotic twins (Table B4).

Next, we treat SAMA as if it was categorical in a multinomial logit model and show that the gender gap is the largest at the top of the mathematics skills' distribution (Table B7).

Lastly, as mentioned in Section 3, we re-estimate Table 1 using three different sets of weights to take selection into the analytical sample into account in the Online Appendix. Table O3 shows that our results stay similar, suggesting that selection to the sample is not a serious concern in this case.

5.2 Sibling peer effects

We now turn our attention to potential peer effects explanations for the gender gap in SAMA. Table 2 confirms our earlier result that having a male co-twin reduces SAMA (Model 1), as we saw before that the gender gap in SAMA is larger among opposite-sex twins. We do not find evidence for a gender-specific relationship because the interaction term of having a male co-twin with female is not significantly different from zero (Model 2).

Our setup does not allow us to test whether the negative effect of having a male co-twin is biological (i.e, stems from in-utero testosterone exposure) or is the result of the different environment into which these young people were born (as opposed to having a same-sex twin). We can however test what happens if we look at the relationship between SAMA and having a brother in general. Note that as mentioned earlier, most siblings in the data are older than the twins. In Model 3, we

control for having a brother (who could be a male twin or a non-twin brother), but we do not find a statistically significant relationship. In Columns 4 and 5, we restrict the sample to opposite-sex twin pairs to look separately at girls with male twins. Repeating Model 3 on this sub-sample (Column 5) does not show a relationship between SAMA and having a brother (on top of one's male co-twin).

Lastly, in Columns 6 and 7, we look at the subsample of same-sex twin pairs (pooling all same-sex twin pairs, girls and boys together). None of the girls in this subsample have a male co-twin. The gender gap among same-sex twins is smaller than the average, 0.27 standard deviation (Column 6), which is consistent with our previous findings showing a larger-than-average gap for opposite-sex twins. Controlling for having a brother and its interaction term with female in Column 7 shows that the gender gap in SAMA is slightly smaller among those who do not have brothers. Although the interaction term of female and having a brother is not statistically significant, it is modest, -0.09 SD. These results are not robust enough to draw a strong conclusion about the role of biological versus environmental factors in the negative association between SAMA and the gender composition of siblings. However, for girls, the relationship between having a male twin versus a non-twin brother and SAMA is similar (-0.086+0.016=-0.07 SD in Column 2 as compared to 0.034-0.094=-0.06 SD in Column 7). For boys, only having a male twin reduces SAMA, having a non-twin brother does not.

Table 2: The role of sibling composition in the gender gap in SAMA

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 1 OS twins	(5) Model 3 OS twins	(6) Model 1 SS twins	(7) Model 3 SS twins
Female	-0.354***	-0.362***	-0.297***	-0.449***	-0.436***	-0.270***	-0.239***
	(0.032)	(0.045)	(0.038)	(0.051)	(0.061)	(0.040)	(0.048)
Has a male twin (MT)	-0.078**	-0.086*					
	(0.032)	(0.045)					
Female*MT		0.016					
		(0.067)					
Has brother			0.042		0.068		0.034
			(0.047)		(0.077)		(0.059)
Female*has brother			-0.084		-0.046		-0.094
			(0.068)		(0.109)		(0.085)
Constant	-0.014	-0.007	-0.076	0.251*	0.256*	-0.193*	-0.207*
	(0.091)	(0.094)	(0.089)	(0.151)	(0.152)	(0.108)	(0.109)
Observations	3,877	3,877	3,877	1,186	1,186	2,691	2,691
R-squared	0.175	0.175	0.174	0.195	0.196	0.171	0.171
Twin FE	No	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample	Total	Total	Total	OS twins	OS twins	SS twins	SS twins

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. 'SS twins' refers to same-sex twins. 'OS twins' refers to opposite-sex twins.

In Table 3, we look at the role of the SAMA of co-twins. On average, own SAMA is positively correlated with co-twin SAMA (Model 1), and this relationship is not different for boys and girls (Model 3). Furthermore, the SAMA of co-twin does not change the previously found negative relationship between having a male co-twin and own SAMA (Model 3). Introducing the triple interaction term of female, having a male co-twin and co-twin SAMA,[4] however, reveals that male co-twin SAMA matters differently for boys and girls (Model 4). For a simpler interpretation, we reestimate Model 4 separately for boys and girls in Columns 5 and 6. For boys (Column 5), SAMA is positively correlated with their male co-twin's SAMA (0.158), while the SAMA of their female co-twin is smaller in magnitude (0.033) and not statistically significant. For girls, it is also true that their SAMA is positively correlated with their same-sex co-twin's SAMA (0.245), however, their SAMA is negatively correlated with their male co-twin's SAMA. In other words, among same-sex male and female twins, high self-assessment is mutually beneficial. Among opposite-sex twins, female SAMA is negatively correlated with male SAMA.[5] Note that this phenomenon does not occur for objective mathematics abilities: the objective mathematics levels of male co-twins do not matter for the gender gap in SAMA (Table O12 in the Online Appendix).

Interestingly, if we repeat the same exercise looking at the gender gap in self-assessed English or physical abilities, we find the same pattern. The confidence of a male co-twin increases the confidence of boys but does not do this for girls in English (Table O14 in the Online Appendix) and physical abilities (Table O15 in the Online Appendix).

Lastly, for easier interpretation, we re-estimate Table 3 using a binary variable capturing very high co-twin SAMA instead of the original continuous variable (Table O13 in the Online Appendix). We create a binary variable for having a "confident twin" that equals one if if the co-twin's SAMA falls in the top 20% of the distribution and zero otherwise. This exercise shows that indeed,

[4]The number of observations in the triple interaction cell is n=598, which is the number of female twins with a male co-twin. The third item of the tripe interaction, the SAMA of co-twin is continuous.

[5]These results are the same for SAMA measured at age 12 after controlling for math test scores (Table B8 in Appendix B), using our IV strategies to correct for measurement error in math levels (Tables B9 and B10 in Appendix B) as well as after reweighting the model with the three types of weights introduced above to account for selection to the analytical sample (Table O5 in the Online Appendix). When we restrict the sample to dizygotic twins in Table B11 in Appendix B, the estimated coefficient on the triple interaction is still negative, but smaller and insignificant. This is most likely the result of decreased sample size.

Table 3: The role of co-twin (CT) SAMA

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.326***	-0.326***	-0.383***	-0.342***		
	(0.030)	(0.031)	(0.035)	(0.033)		
Has a male twin (MT)			-0.151***	-0.127***	-0.127***	-0.122***
			(0.036)	(0.033)	(0.044)	(0.047)
SAMA of CT, age 9, std	0.153***	0.127***	0.181***	0.034	0.033	0.245***
	(0.023)	(0.029)	(0.028)	(0.036)	(0.036)	(0.036)
MT*SAMA of CT			-0.034	0.162***	0.158***	-0.206***
			(0.039)	(0.054)	(0.055)	(0.057)
Female*SAMA of CT		0.048		0.207***		
		(0.037)		(0.050)		
Female*MT*SAMA of CT				-0.365***		
				(0.094)		
Constant	-0.047	-0.046	0.047	0.025	-0.019	-0.272**
	(0.081)	(0.081)	(0.085)	(0.083)	(0.114)	(0.109)
Observations	3,722	3,722	3,722	3,722	1,707	2,015
R-squared	0.196	0.197	0.201	0.208	0.205	0.158
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. CT refers to co-twins.

among boys, having a confident male twin comes with higher SAMA (-0.158+0.127+0.234=0.203), but this is not the case among girls (0.002+0.403-0.485=-0.08).

5.3 The transmission of gender stereotypes

The transmission of gender stereotypes from adults to children may be an important driver of the gender gap in SAMA. Unfortunately, we are unable to identify the causal effects of parental assessments on their children's assessments, as these two can mutually enforce each other and thus are endogenous. However, we explore whether there is a gender gap in how parents and teachers assess the mathematics ability of boys and girls. After finding gender differences in these assessments, we then control for them in our main model.

In Tables O9 and O10 in the Online Appendix, we estimate the same models as in Table 1, but

now the outcome variable is either parent or teacher assessment of the twins mathematics ability instead of SAMA. The main results are broadly similar. Parents assess girls' mathematics ability lower than boys even once we account for their actual mathematics performance (Model 2, approximately -0.2 SD). Interestingly, the difference is even more pronounced between boys and girls within the same twin pair (Model 5). Here parents assess their daughters' mathematics ability -0.42 SD lower than their male twins.

The gender gap in teachers' assessment of boys' and girls' mathematics ability is similar in magnitude to parents' assessment in raw terms (-0.2 SD), but halves once we account for actual mathematics ability, i.e. mathematics levels given by the same teachers (-0.12 SD).[6] Teachers should have more accurate knowledge about the children's actual mathematics ability, so this is unsurprising. Including twin fixed effects in the model does not change the estimated coefficient significantly. Next, we explore the inclusion of parent and teacher assessments as a potential channel by including them in our main models of SAMA.

[6]Theoretically, teachers could also show a gender bias when they determine the mathematics levels of kids. We tested on the age 12 sample whether there is a gender gap in mathematics levels. Interestingly, while there is a raw gender gap in mathematics levels, once mathematics test scores and cognitive abilities are controlled for, this gap becomes small and non-significant. While we cannot test the same thing on the age nine sample as test scores are only available for age 12, we believe that these results would be similar.

Table 4: The role of parental and teachers assessments in the gender gap in SAMA

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.227***	-0.079	-0.288***	-0.099	-0.217***	-0.004
	(0.029)	(0.141)	(0.031)	(0.132)	(0.029)	(0.147)
Parental assessment of Math	0.498***	0.518***			0.465***	0.463***
	(0.021)	(0.028)			(0.022)	(0.031)
Female*parental assessment		-0.037				0.005
		(0.033)				(0.040)
Teachers' assessment of Math			0.355***	0.384***	0.161***	0.197***
			(0.029)	(0.033)	(0.028)	(0.033)
Female*teachers' assessment				-0.056		-0.069*
				(0.036)		(0.041)
Constant	-1.908***	-1.989***	-1.237***	-1.331***	-2.321***	-2.429***
	(0.112)	(0.134)	(0.128)	(0.139)	(0.131)	(0.148)
Observations	3,877	3,877	3,877	3,877	3,877	3,877
R-squared	0.308	0.309	0.207	0.207	0.315	0.315
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

The models in Table 4 highlight the importance of parental perceptions in explaining the gender gap. Model 1 shows a decrease of approximately 30 percent when we introduce parental assessments of their children's mathematics ability (from -0.32 SD in Table 1 to -0.23 SD). The coefficient on the interaction term of female with parental assessment in Model 2 is not significant. This means that parental assessment in general does not have a different correlation with the SAMA of boys and girls.

Compared to the main model (Model 3 in Table 1), the gap is also reduced somewhat when we control for teacher assessments in Column 3, but not by as much. Again, the interaction term of teachers' assessments and female is not significant (Column 4). Introducing both sets of adult assessments in Model 5 reduces the gap slightly more, but it seems as though most of the reduction is led by parental assessments. Introducing the interaction terms of parental and teachers' assessments with gender reveals that conditional on parental assessment, girls' SAMA is negatively correlated with teachers' assessments (Model 6). The same models estimated with twin pair/family FEs are shown in Table O11 in the Online Appendix.

Table 5: The role of stereotypically gender-biased parental assessments in the gender gap in SAMA

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.324***	-0.324***	-0.032	-0.447***	-0.447***	-0.146**
	(0.032)	(0.032)	(0.037)	(0.051)	(0.051)	(0.061)
Stereotypically assessed person		0.002	0.576***		0.027	0.512***
		(0.034)	(0.043)		(0.051)	(0.076)
Female*stereotypically assessed			-1.093***			-0.961***
			(0.067)			(0.116)
Constant	-0.064	-0.065	-0.139	0.033	0.027	-0.090
	(0.089)	(0.089)	(0.085)	(0.236)	(0.237)	(0.230)
Observations	3,877	3,877	3,877	3,877	3,877	3,877
R-squared	0.174	0.174	0.227	0.164	0.164	0.199
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics levels at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. The measure of parental stereotypical assessment was created the following way. First, a multinomial logit model of the form $f(k,i) = beta_k * x_i$ is estimated, where $beta_k$ is a set of regression coefficients associated with the k terciles of parental assessments, $k = 1, 2, 3$, and x_i is the set of three explanatory variables: mathematics levels and verbal and non-verbal cognitive skills at age nine. Then, predicted categories of parental assessment are fitted by the model and they are compared to the observed parental assessment of individuals. An individual is over(under)estimated if their observed parental assessment category is higher(lower) than their predicted category.

We introduce our measure of stereotypically gender-biased parental assessment as explained above[7] in Table 5. Compared to the gender gap in our main model (-0.32 SD, Model 1 in Table 1), the gap does not change when we introduce the measure in Model 2 (-0.32 SD). However, when we also introduce its interaction term with gender (Model 3), the gender gap among those who did not receive a stereotypically biased parental assessment becomes small and insignificant (-0.03 SD). This shows that the average difference between the SAMA of non-underestimated girls and non-overestimated boys is statistically negligible. The coefficient on the stereotypical assessment measure is positive and significant (0.58 SD), indicating that the average SAMA of overestimated boys is larger than that of non-overestimated boys. Lastly, the estimated coefficient on the interaction term is large and highly significant (-1.1 SD). This suggests that the average SAMA of underestimated girls is more than one standard deviation lower than the SAMA of overestimated boys. Results are similar in the twin FE setup (Models 4-6), when we use our alternative measure of parental stereotypical assessments (Table B12 in Appendix B), and also when we apply our IV strategies for measurement error (Tables B13 and B13 in Appendix B).[8]

Reweighting these models with the three types of weights introduced above also leads to similar conclusions (Table O4 in the Online Appendix). These results suggest that gender-biased parental assessments play a large role in the gender gap in SAMA.[9]

[7]The measure of gender-stereotypical parental assessment is a binary variable that captures whether parents' assessment of their children's mathematics abilities is stereotypically gender-biased, i.e. they overestimate their son in mathematics and/or they underestimate their daughter in math.

[8]The fact that stereotypical parental assessments are associated with SAMA raises the question how they might impact within-twin peer effects (that we explore in Table 3). Table O18 in the Online Appendix investigates this question. We split the samples boys and girls to subsamples of stereotypically assessed individuals (i.e., overestimated boys and underestimated girls) and not-stereotypically assessed individuals (not overestimated boys and not underestimated girls), resulting in four subsamples altogether. Interestingly, among girls, it does not matter whether they are stereotypically assessed by their parents or not: the large negative correlation between their SAMA and the SAMA of their male co-twin is the same in the two female subsamples (Columns (3) and (4)). Among boys, however, the positive correlation between their own SAMA and the SAMA of their co-twin is only there among overestimated boys. This suggests that the relationship we find for girls is probably more society-driven, while the association for boys is more family-driven.

[9]Ideally, we would also want to look at the role of gender roles in the home using alternative measures. Unfortunately, the data do not include direct measures of gender roles. Interestingly, SAMA is neither correlated with parental education (Table O16 in the Online Appendix) nor with the characteristics of maternal employment (Table O17 in the Online Appendix). We have also tried to determine whether the relative educational or employment characteristics of mothers matter (i.e. if they have higher educational attainment or work in higher-status jobs than fathers), but they do not. We believe that the stereotypical assessment of their children's mathematics skills is the best measure of parental gender stereotypes in TEDS.

5.4 The role of girls' comparative advantage in English

Table 6 investigates the role of girls' comparative advantage in English in the gender gap in SAMA. Extending our main models, Model 3 and Model 4 in Table 1, by controlling for English ability (levels) slightly decreases the gap by about 5-10 percent (from 0.32 SD to 0.28 SD in the OLS model and from 0.45 SD to 0.43 in the FE model). The coefficients on English levels are statistically significant and negative: those with better English skills have lower confidence in their mathematics skills, conditional on their mathematics abilities. Interestingly, when self-assessed English ability is also added to the model in Column 2 and 5, the gender gap bounces back to the earlier levels. SAMA is positively correlated with English self-assessment in all models. The interaction terms of female are negative with English levels and positive with English self-assessment, but the former is only significant in the FE model (Column 6) while the latter is only significant in the OLS model (Column 3). Thus, in general, the positive correlation between confidence in English and confidence in mathematics is about 30 percent larger for girls than for boys. Among opposite-sex twins, however, English levels seem to matter more for girls: English levels decrease girls' SAMA two times as much as boys' confidence in math.

Table 6: The role of girls' comparative advantage in English in the gender gap in SAMA

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 2	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.283***	-0.318***	-0.734***	-0.425***	-0.462***	-0.798***
	(0.033)	(0.031)	(0.211)	(0.053)	(0.052)	(0.294)
English level, age 9	-0.116***	-0.214***	-0.186***	-0.067	-0.129***	-0.083*
	(0.024)	(0.023)	(0.028)	(0.043)	(0.043)	(0.049)
Perceived English, age 9		0.387***	0.334***		0.229***	0.185***
		(0.026)	(0.035)		(0.036)	(0.050)
Female*English level			-0.050			-0.089*
			(0.033)			(0.052)
Female*Self-assessed English			0.102**			0.085
			(0.050)			(0.069)
Constant	-0.087	-1.664***	-1.445***	0.011	-0.897***	-0.739**
	(0.089)	(0.134)	(0.165)	(0.236)	(0.272)	(0.307)
Observations	3,877	3,877	3,877	3,877	3,877	3,877
R-squared	0.179	0.242	0.243	0.165	0.188	0.189
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

6 Discussion

This paper examined the gender gap in self-assessed mathematics ability using rich data on twins born in the UK. Despite a range of literature on the gender gap in mathematics performance and STEM attainment more broadly, literature exploring the gender gap in the self-assessment of mathematics ability is limited. We set out to fill this gap and examine why boys are more likely to rate their mathematics ability higher than girls, even when their ability is the same.

We find that the gender gap in SAMA is about three times as large as the gender gap in objective mathematics ability. Objective skills only explain 14-26% of the gender gap in SAMA. Interestingly, the gender gap in SAMA is even larger among opposite-sex twins than among non-related boys and girls. We probe these results further and explore three potential channels: sibling peer

effects, the transmission of gendered stereotypes from adults to children, and girls' comparative advantage in English.

In terms of twin peer effects, we find that the SAMA of boys is positively correlated with the SAMA of a male twin, but this positive correlation is not present for girls. This supports the idea that within families, there might be a narrative of who is the "mathematics person" and who is not. Once this role has been taken (by the male twin), it is difficult for the female twin to view herself as a "mathematics person" as well. While the SAMA of one's co-twin is undoubtedly endogenous, these results might highlight the role of environment in terms of growing up with a male sibling as one's most direct point of comparison. Psychologists point to social comparison theory (Festinger, 1954) and contrast effects (Morse and Gergen, 1970) to describe how individuals shape their self-perceptions based on others, which falls under the umbrella of peer effects in the economics literature (Sacerdote, 2011).

Interestingly, the objective mathematics ability of the co-twin does not matter for either boys or girls, only their self-assessment. This again points to the importance of stereotypes pervading sibling interactions as opposed to actual ability. The peer effects literature in economics has also highlighted the importance of non-cognitive peer effects over and above traditional cognitive peer effects (Golsteyn et al. (2021); Shure (2021)), which is in line with our finding.

We also find that the confidence of a male twin works the same way for self-assessed English and physical abilities as for SAMA: the confidence of a male twin is positively correlated with the confidence of boys but not with the confidence of girls even in English, where girls are better on average at than boys. While again, our results are not causal, they might offer a potential explanation for the gender gap in labor market outcomes, especially in top jobs and high-level managerial positions. For women, exposure to highly confident men might be more difficult than for men. As top job positions are traditionally filled by confident men, women might suffer a double penalty: not only are they less confident than men, as shown by Adamecz-Völgyi and Shure (2022), but their confidence is not supported in those environments (while men's confidence might be). This phenomenon may serve as a barrier to both entry and progression for women in top jobs.

Our results are in line with the literature on the transmission of gendered stereotypes from adults to the next generation. Parental assessments of the mathematics performance of their children (conditional on objective skills) explain a further 23% of the gender gap in SAMA. Furthermore, we find that most of the gender gap is driven by families where parents assess their children according to gender stereotypes, i.e. assess boys higher and girls lower in mathematics. For those children in families without stereotypical assessments, the gender gap in self-assessments is small. Unfortunately, teachers are not immune to this and also over-assess boys and under-assess girls; however, this explains a smaller portion of the gender gap in SAMA. We cannot exclude, however, two potential sources of endogeneity between parents' and children's assessments. First, parents might have some unobserved knowledge about the math abilities of their children, that is above and beyond their objective math levels and their teachers' assessments, hence they play such an important role in terms of explaining the gender gap in SAMA. Second, parental assessments might mirror the kids' own assessment, hence they are so highly correlated. Identifying the causal effects of parental assessments/stereotypes on their children's own assessments is an extremely challenging exercise that has not been solved yet.

Although we find that girls have a comparative advantage in English, this does not explain the gender gap in SAMA. Having higher English ability or higher self-assessed English ability does not reduce the gender gap in SAMA. Girls are not specializing in one domain at the expense of another.

There are potential explanations behind our findings that could not be explored in this paper. This includes in-utero testosterone exposure (Auyeung et al. (2009);Gielen et al. (2016)). There is a strand of literature that looks at the effects of in-utero testosterone exposure and shows that those with a male co-twin in-utero have different life outcomes than those with a female co-twin (or no twin sibling), even if their twin brother passed away shortly after birth (Bütikofer et al., 2019). We are unfortunately unable to probe this further with our data, but it would support our results that the gender gap is larger in opposite-sex twin pairs.

Our study also has some caveats. First, unobserved facets of mathematics ability, which might

be known by kids/parents/teachers, but not measured by mathematics levels or cognitive skills could hinder our results. Despite our efforts to carry out various robustness checks around our measures of mathematics ability, they may be subject to some degree of measurement error. Reassuringly, when we replicated our main result using the age 12 data that also captured math test scores besides math levels, we received similar results. However, parental assessments are not available at age 12, so we could not test our results on the role of parental assessments. Second, parents, teachers, and the twins were all asked to assess their mathematics ability in the same wave. As mentioned above, it may be the case that children's self-assessments shape their parents' or teachers' assessments as much as the adults' assessments shape the children's. We unfortunately cannot account for the direction of this relationship since the parents and teachers were only asked about the twins' mathematics ability in one wave. Third, as again mentioned above, while the gender of a co-twin is random, their confidence is not. Lab experiments are needed to test what happens to the gender gap in confidence when women/men are randomly exposed to more confident male/female peers.

In terms of policy, our results suggest that potential interventions to reduce the gender gap in SAMA should also target parents and teachers, not just children. It is not enough to inspire girls into STEM fields, systematic change around who adults frame as the "mathematics person" is also needed. Teacher training could include further emphasis on unconscious bias in marking and assessment. Parents should be aware of the narratives they develop within families to place children into "math" or "verbal" person categories as this early differentiation can have long-lasting consequences (Chaffee and Plante, 2022).

References

Adamecz-Völgyi, A., Shure, N., 2022. The gender gap in top jobs – the role of overconfidence. Labour Economics , 102283URL: https://linkinghub.elsevier.com/retrieve/pii/S0927537122001737, doi:10.1016/j.labeco.2022.102283.

Alicke, M.D., Dunning, D.A., Krueger, J., 2005. The Self in Social Judgment. Psychology Press. Google-Books-ID: hEoG8OrIR7sC.

Auyeung, B., Baron-Cohen, S., Ashwin, E., Knickmeyer, R., Taylor, K., Hackett, G., Hines, M., 2009. Fetal Testosterone Predicts Sexually Differentiated Childhood Behavior in Girls and in

Boys. Psychological Science 20, 144–148. URL: http://journals.sagepub.com/doi/10.1111/j.1467-9280.2009.02279.x, doi:10.1111/j.1467-9280.2009.02279.x.

Baird, C.L., Keene, J.R., 2019. Closing the Gender Gap in Math Confidence: Gender and Race/Ethnic Similarities and Differences. International Journal of Gender, Science and Technology 10, 33. URL: http://genderandset.open.ac.uk/index.php/genderandset/article/view/452. edition: 2019-02-11 ISBN: 2040-0748 Type: gender; confidence; math; STEM; race/ethnicity.

Bhalotra, S., Clarke, D., 2019. Twin Birth and Maternal Condition. The Review of Economics and Statistics 101, 853–864. URL: https://direct.mit.edu/rest/article/101/5/853/58541/Twin-Birth-and-Maternal-Condition, doi:10.1162/rest_a_00789.

Bharadwaj, P., De Giorgi, G., Hansen, D., Neilson, C.A., 2016. The Gender Gap in Mathematics: Evidence from Chile. Economic Development and Cultural Change 65, 141–166. URL: https://www.journals.uchicago.edu/doi/10.1086/687983, doi:10.1086/687983.

Breda, T., Napp, C., 2019. Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. Proceedings of the National Academy of Sciences 116, 15435–15440. URL: https://www.pnas.org/doi/10.1073/pnas.1905779116, doi:10.1073/pnas.1905779116. publisher: Proceedings of the National Academy of Sciences.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. URL: https://link.springer.com/article/10.1023/A:1010933404324, doi:10.1023/A:1010933404324.

Bütikofer, A., Figlio, D.N., Karbownik, K., Kuzawa, C.W., Salvanes, K.G., 2019. Evidence that prenatal testosterone transfer from male twins reduces the fertility and socioeconomic success of their female co-twins. Proceedings of the National Academy of Sciences 116, 6749–6753. URL: https://www.pnas.org/doi/10.1073/pnas.1812786116, doi:10.1073/pnas.1812786116. publisher: Proceedings of the National Academy of Sciences.

Chaffee, K.E., Plante, I., 2022. How Parents' Stereotypical Beliefs Relate to Students' Motivation and Career Aspirations in Mathematics and Language Arts. Frontiers in Psychology 12. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.796073.

Collischon, M., Eberl, A., 2020. Let's Talk About Fixed Effects: Let's Talk About All the Good Things and the Bad Things. KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 72, 289–299. URL: https://doi.org/10.1007/s11577-020-00699-8, doi:10.1007/s11577-020-00699-8.

Dossi, G., Figlio, D., Giuliano, P., Sapienza, P., 2021. Born in the family: Preferences for boys and the gender gap in math. Journal of Economic Behavior & Organization 183, 175–188. URL: https://www.sciencedirect.com/science/article/pii/S0167268120304716, doi:10.1016/j.jebo.2020.12.012.

Dunning, D., Heath, C., Suls, J.M., 2004. Flawed Self-Assessment: Implications for Health, Education, and the Workplace. Psychological Science in the Public Interest 5, 69–106. URL: http://journals.sagepub.com/doi/10.1111/j.1529-1006.2004.00018.x, doi:10.1111/j.1529-1006.2004.00018.x.

Eagly, A.H., Wood, W., 2012. Social Role Theory, in: Handbook of Theories of Social Psychology. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, pp. 458–476. URL: http://sk.sagepub.com/reference/hdbk_socialpsychtheories2/n49.xml, doi:10.4135/9781446249222.n49.

Ehrlinger, J., Dunning, D., 2003. How chronic self-views influence (and potentially mislead) estimates of performance. Journal of Personality and Social Psychology 84, 5–17. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.84.1.5, doi:10.1037/0022-3514.84.1.5.

Festinger, L., 1954. A Theory of Social Comparison Processes. Human Relations 7, 117–140. URL: http://journals.sagepub.com/doi/10.1177/001872675400700202, doi:10.1177/001872675400700202.

Friedman, J., Hastie, T., Tibshirani, R., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second edition ed., Springer.

Fryer, R.G., Levitt, S.D., 2010. An Empirical Analysis of the Gender Gap in Mathematics. American Economic Journal: Applied Economics 2, 210–240. URL: https://pubs.aeaweb.org/doi/10.1257/app.2.2.210, doi:10.1257/app.2.2.210.

Gielen, A.C., Holmes, J., Myers, C., 2016. Prenatal Testosterone and the Earnings of Men and Women. Journal of Human Resources 51, 30–61. URL: http://jhr.uwpress.org/cgi/doi/10.3368/jhr.51.1.30, doi:10.3368/jhr.51.1.30.

Gillen, B., Snowberg, E., Yariv, L., 2019. Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. Journal of Political Economy 127, 1826–1863. URL: https://www.journals.uchicago.edu/doi/full/10.1086/701681, doi:10.1086/701681. publisher: The University of Chicago Press.

Golsteyn, B.H., Non, A., Zölitz, U., 2021. The impact of peer personality on academic achievement. Journal of Political Economy 129, 1052–1099. ISBN: 0022-3808 Publisher: The University of Chicago Press Chicago, IL.

Goulas, S., Griselda, S., Megalokonomou, R., 2020. Comparative Advantage and Gender Gap in Stem. SSRN Electronic Journal URL: https://www.ssrn.com/abstract=3620627, doi:10.2139/ssrn.3620627.

Grömping, U., 2009. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician 63, 308–319. URL: https://doi.org/10.1198/tast.2009.08199, doi:10.1198/tast.2009.08199.

Hainmueller, J., 2012. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. Political Analysis 20, 25–46. URL: https://www.cambridge.org/core/product/identifier/S1047198700012997/type/journal_article, doi:10.1093/pan/mpr025.

Kay, K., Shipman, C., 2014. The Confidence Code: The Science and Art of Self-Assurance—What Women Should Know. 1st edition ed., Harper Business, New York, NY.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer Science & Business Media. Google-Books-ID: xYRDAAAAQBAJ.

Ladd, H.F., Walsh, R.P., 2002. Implementing value-added measures of school effectiveness: getting the incentives right. Economics of Education Review 21, 1–17. URL: https://www.sciencedirect.com/science/article/pii/S027277570000039X, doi:10.1016/S0272-7757(00)00039-X.

Lavy, V., Schlosser, A., 2011. Mechanisms and Impacts of Gender Peer Effects at School. American Economic Journal: Applied Economics 3, 1–33. URL: https://www.aeaweb.org/articles?id=10.1257/app.3.2.1, doi:10.1257/app.3.2.1.

Makarova, E., Aeschlimann, B., Herzog, W., 2019. The Gender Gap in STEM Fields: The Impact of the Gender Stereotype of Math and Science on Secondary Students' Career Aspirations. Frontiers in Education 4, 60. URL: https://www.frontiersin.org/article/10.3389/feduc.2019.00060/full, doi:10.3389/feduc.2019.00060.

Marsh, H.W., Hau, K.T., 2004. Explaining Paradoxical Relations Between Academic Self-Concepts and Achievements: Cross-Cultural Generalizability of the Internal/External Frame of Reference Predictions Across 26 Countries. Journal of Educational Psychology 96, 56–67. doi:10.1037/0022-0663.96.1.56. place: US Publisher: American Psychological Association.

Morse, S., Gergen, K.J., 1970. Social comparison, self-consistency, and the concept of self. Journal of Personality and Social Psychology 16, 148–156. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/h0029862, doi:10.1037/h0029862.

Mullis, I.V.S., Martin, M.O., Foy, P., Kelly, D.L., Fishbein, B., 2020. TIMSS 2019 International Results in Mathematics and Science. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/. Technical Report.

Nicoletti, C., Rabe, B., 2019. Sibling spillover effects in school achievement. Journal of Applied Econometrics 34, 482–501. URL: https://onlinelibrary.wiley.com/doi/10.1002/jae.2674, doi:10.1002/jae.2674.

Nicoletti, C., Sevilla, A., Tonei, V., 2022. Gender Stereotypes in the Family. SSRN Electronic Journal URL: https://www.ssrn.com/abstract=4294398, doi:10.2139/ssrn.4294398.

Niederle, M., Vesterlund, L., 2007. Do Women Shy Away From Competition? Do Men Compete Too Much? The Quarterly Journal of Economics 122, 1067–1101. URL: https://academic.oup.com/qje/article/122/3/1067/1879500, doi:10.1162/qjec.122.3.1067.

OECD, 2020. PISA 2018 Results (Volume VI): Are Students Ready to Thrive in an Interconnected World? PISA, OECD. URL: https://www.oecd-ilibrary.org/education/pisa-2018-results-volume-vi_d5f68679-en, doi:10.1787/d5f68679-en.

Rimfeld, K., Malanchini, M., Spargo, T., Spickernell, G., Selzam, S., McMillan, A., Dale, P.S., Eley, T.C., Plomin, R., 2019. Twins Early Development Study: A Genetically Sensitive Investigation into Behavioral and Cognitive Development from Infancy to Emerging Adulthood. Twin Research and Human Genetics 22, 508–513. URL: https://www.cambridge.org/core/product/identifier/S1832427419000562/type/journal_article, doi:10.1017/thg.2019.56.

Sacerdote, B., 2011. Peer effects in education: How might they work, how big are they and how much do we know thus far?, in: Handbook of the Economics of Education. Elsevier. volume 3, pp. 249–277.

Shure, N., 2021. Non-cognitive peer effects in secondary education. Labour Economics 73, 102074. URL: https://linkinghub.elsevier.com/retrieve/pii/S0927537121001093, doi:10.1016/j.labeco.2021.102074.

Sterling, A.D., Thompson, M.E., Wang, S., Kusimo, A., Gilmartin, S., Sheppard, S., 2020. The confidence gap predicts the gender pay gap among STEM graduates. Proceedings of the National Academy of Sciences 117, 30303–30308. URL: https://pnas.org/doi/full/10.1073/pnas.2010269117, doi:10.1073/pnas.2010269117.

Walker, I., Zhu, Y., 2011. Differences by degree: Evidence of the net financial rates of return to undergraduate study for England and Wales. Economics of Education Review 30, 1177–1186. URL: https://linkinghub.elsevier.com/retrieve/pii/S0272775711000033, doi:10.1016/j.econedurev.2011.01.002.

Appendix

A Descriptive statistics

Table A1: Descriptive statistics, age nine sample

	Mean	SD	Min	Max	N
Female	0.54	0.50	0.00	1.00	3,877
Cohort born between Jan 94-Aug 94	0.38	0.48	0.00	1.00	3,877
Cohort born between Sep 94-Aug 95	0.62	0.48	0.00	1.00	3,877
Self-assessed Math (SAMA), age 9	3.83	0.99	1.00	5.00	3,877
Math level, age 9	0.09	0.97	-2.94	2.99	3,877
Verbal abilities, age 9	0.06	0.96	-3.34	2.61	3,877
Non-verbal abilities, age 9	0.07	0.96	-3.72	1.39	3,877
Elder twin	0.50	0.50	0.00	1.00	3,877
Heavier twin at birth	0.47	0.50	0.00	1.00	3,877
Birthweight, gramms	2,536.82	546.79	595.88	6,320.00	3,877
Verbal abilities, age 7	0.10	0.98	-3.04	5.90	3,473
Non-verbal abilities, age 7	0.08	0.96	-3.64	2.53	3,487
Math level, age 7	0.05	0.93	-3.68	3.23	3,059
Self-assessed English (SAEA), age 9	4.11	0.70	1.00	5.00	3,877
English level, age 9	0.10	0.96	-3.08	3.07	3,877
Parental assessment of Math	3.94	0.93	1.00	5.00	3,877
Teachers' assessment of Math	3.37	0.83	1.00	5.00	3,877
Has a male twin (MT)	0.46	0.50	0.00	1.00	3,877
Has brother	0.32	0.47	0.00	1.00	3,877
Overestimated in Math	0.23	0.42	0.00	1.00	3,877
Underestimated in Math	0.23	0.42	0.00	1.00	3,877
Stereotipically assessed person	0.26	0.44	0.00	1.00	3,877
Stereotipically assessed person, 5 cat	0.32	0.47	0.00	1.00	3,877
No qual or low-grade CSE/GCSE	0.11	0.31	0.00	1.00	3,863
High-grade CSE/GCSE	0.31	0.46	0.00	1.00	3,863
A-level or below degree	0.28	0.45	0.00	1.00	3,863
Degree	0.31	0.46	0.00	1.00	3,863
Mother has A-levels or above	0.42	0.49	0.00	1.00	3,877
Mother has managerial job	0.11	0.31	0.00	1.00	3,877
Mother needs qualification	0.25	0.44	0.00	1.00	3,877
SAMA of CT, age 9	3.82	0.99	1.00	5.00	3,877
Math level of CT, age 9	0.08	0.97	-2.94	2.99	3,877
Self-assessed physical (SAPA), age 9	4.44	0.67	1.00	5.00	3,867
SAEA of CT, age 9	0.04	0.97	-4.28	1.29	3,876
SAPA of CT, age 9	0.02	0.98	-4.97	0.84	3,863

Source: TEDS (Rimfeld et al., 2019).

Table A2: The gender gap in our main measures, age nine sample

	Boys	Girls	Gap	SE	P-values	Obs
Verbal abilities, age 7	0.105	0.097	-0.009	0.037	0.815	3473
Non-verbal abilities, age 7	0.079	0.080	0.001	0.036	0.976	3487
Math level, age 7	0.078	0.023	-0.056	0.038	0.148	3059
Verbal abilities, age 9	0.083	0.037	-0.045	0.035	0.194	3877
Non-verbal abilities, age 9	0.058	0.079	0.021	0.035	0.539	3877
Math level, age 9	0.157	0.029	-0.129	0.035	0.000	3877
Self-assessed Math (SAMA), age 9, std	0.204	-0.173	-0.377	0.034	0.000	3877
Self-assessed English (SAEA), age 9, std	-0.121	0.103	0.224	0.035	0.000	3877
Self-assessed physical (SAPA), age 9	0.069	-0.014	-0.083	0.034	0.014	3867
Parental assessment of Math, std	0.157	-0.134	-0.291	0.037	0.000	3877
Teachers' assessment of Math, std	0.121	-0.103	-0.224	0.036	0.000	3877
Verbal abilities, age 12	0.372	0.108	-0.263	0.042	0.000	2469
Non-verbal abilities, age 12	0.337	0.286	-0.052	0.042	0.219	2398
SAMA, age 12, std	0.324	-0.081	-0.405	0.089	0.000	507
Math level, age 12	0.506	0.456	-0.050	0.051	0.331	1497
Math test scores, age 12, std	0.485	0.213	-0.272	0.087	0.002	507
Overestimated in Math	0.266	0.209	-0.057	0.015	0.000	3877
Underestimated in Math	0.195	0.264	0.069	0.015	0.000	3877
Stereotipically assessed person	0.266	0.264	-0.002	0.016	0.915	3877
Stereotipically assessed person, 5 cat	0.230	0.394	0.163	0.017	0.000	3877

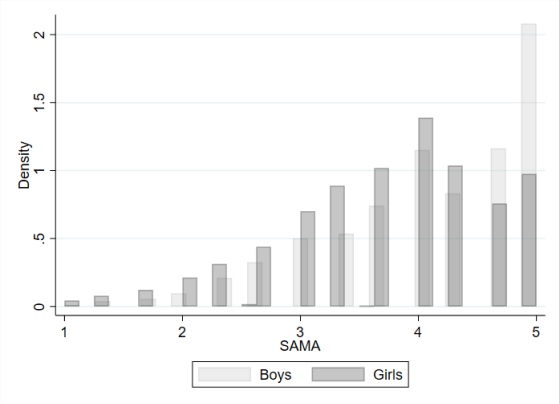
Source: TEDS (Rimfeld et al., 2019).

Table A3: Correlation matrix of measures, age nine

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) SAMA	1	0.22	0.37	0.17	0.29	0.20	0.19	0.20	0.54	0.42
(2) SAMA of co-twin	0.22	1	0.17	0.37	0.15	0.12	0.11	0.10	0.20	0.19
(3) Math level	0.37	0.17	1	0.53	0.23	0.72	0.33	0.38	0.58	0.78
(4) Math level of co-twin	0.17	0.37	0.53	1	0.13	0.47	0.23	0.26	0.31	0.44
(5) Self-assessed English	0.29	0.15	0.23	0.13	1	0.35	0.19	0.11	0.20	0.22
(6) English levels	0.20	0.12	0.72	0.47	0.35	1	0.34	0.33	0.41	0.59
(7) Verbal abilities	0.19	0.11	0.33	0.23	0.19	0.34	1	0.40	0.29	0.30
(8) Non-verbal abilities	0.20	0.10	0.38	0.26	0.11	0.33	0.40	1	0.33	0.36
(9) Parental assessment	0.54	0.20	0.58	0.31	0.20	0.41	0.29	0.33	1	0.61
(10) Teachers assessment	0.42	0.19	0.78	0.44	0.22	0.59	0.30	0.36	0.61	1

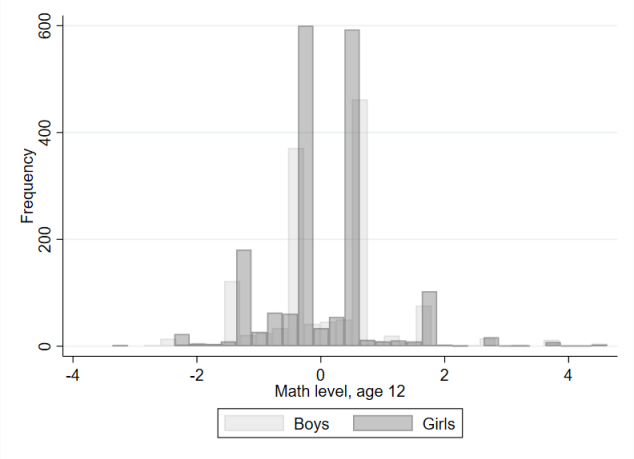
Source: TEDS (Rimfeld et al., 2019). Number of observations: 3,877.

Figure A1: The distribution of mathematics self-assessment, age 12



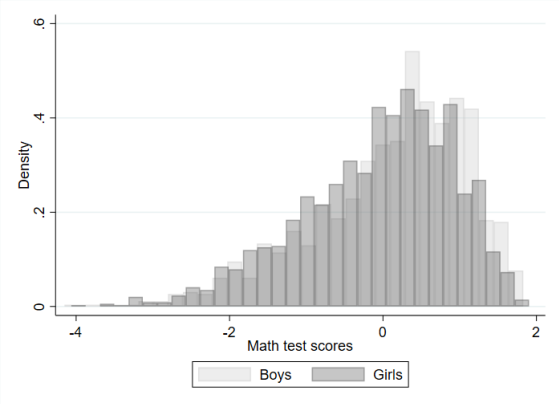
Source: TEDS (Rimfeld et al., 2019). Number of observations: 3,196.

Figure A2: The distribution of mathematics levels, age 12



Source: TEDS (Rimfeld et al., 2019). Number of observations: 3,196.

Figure A3: The distribution of mathematics test scores, age 12



Source: TEDS (Rimfeld et al., 2019). Number of observations: 3,196.

B Robustness checks

Table B1: The gender gap in SAMA at age nine - controlling for math levels and abilities from age seven

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.328***	-0.310***	-0.307***	-0.449***	-0.506***	-0.491***
	(0.032)	(0.036)	(0.036)	(0.052)	(0.060)	(0.059)
Math level, age 9		0.333***	0.284***		0.363***	0.309***
		(0.021)	(0.025)		(0.038)	(0.041)
Math level, age 7			0.132***			0.197***
			(0.025)			(0.044)
Verbal abilities, age 9	0.063***	0.062***	0.062***	0.095***	0.114***	0.096**
	(0.018)	(0.021)	(0.022)	(0.036)	(0.043)	(0.043)
Non-verbal abilities, age 9	0.070***	0.070***	0.055**	0.148***	0.130***	0.107***
	(0.020)	(0.023)	(0.023)	(0.033)	(0.040)	(0.039)
Verbal abilities, age 7			-0.067***			-0.084**
			(0.021)			(0.036)
Non-verbal abilities, age 7			0.024			0.058*
			(0.020)			(0.031)
Math level = 2	0.443***			0.446***		
	(0.045)			(0.069)		
Math level = 3	0.824***			0.888***		
	(0.050)			(0.085)		
Constant	-0.499***	-0.112	-0.091	-0.416*	0.157	0.112
	(0.094)	(0.102)	(0.102)	(0.243)	(0.274)	(0.272)
Observations	3,877	2,942	2,942	3,877	2,942	2,942
R-squared	0.164	0.175	0.186	0.156	0.177	0.195
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: elder twin, heavier twin, and birth weight.

Table B2: The gender gap in SAMA at age nine - using age seven math levels as an IV for age nine math levels

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 3 OS subsample	(5) Model 4
Female	-0.352***	-0.297***	-0.297***	-0.454***	-0.457***
	(0.039)	(0.037)	(0.037)	(0.059)	(0.063)
Math level, age 9		0.549***	0.475***	0.499***	0.717***
		(0.035)	(0.043)	(0.073)	(0.110)
Verbal abilities, age 9			0.032	-0.060	0.040
			(0.023)	(0.037)	(0.049)
Non-verbal abilities, age 9			0.027	0.043	0.040
			(0.026)	(0.041)	(0.048)
Constant	0.174***	0.055	-0.089	0.197	0.108
	(0.040)	(0.038)	(0.103)	(0.166)	(0.280)
Observations	2,942	2,942	2,942	901	2,942
R-squared	0.031	0.139	0.160	0.189	
Twin FE	No	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	Yes	No

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: elder twin, heavier twin, and birth weight. 2SLS estimates using age seven math levels as instrumental variables for age nine math levels.

Table B3: The gender gap in SAMA at age nine - using the ORIV approach of Gillen et al. (2019)

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 3 OS subsample	(5) Model 4
Female	-0.347***	-0.298***	-0.291***	-0.413***	-0.604***
	(0.039)	(0.036)	(0.036)	(0.058)	(0.075)
Math levels		0.574***	0.568***	0.583***	-0.601***
		(0.030)	(0.039)	(0.061)	(0.097)
Verbal abilities, age 9			0.018	-0.088**	0.313***
			(0.022)	(0.036)	(0.056)
Non-verbal abilities, age 9			-0.012	0.000	0.354***
			(0.025)	(0.040)	(0.053)
Constant	0.165***	0.129***	0.006	0.193	0.208
	(0.039)	(0.037)	(0.102)	(0.167)	(0.338)
Observations	6,118	6,118	6,118	1,874	6,118
R-squared	0.031	0.081	0.085	0.117	
Twin FE	No	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	Yes	No

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: elder twin, heavier twin, and birth weight. 2SLS estimates using the ORIV approach of Gillen et al. (2019).

Table B4: The gender gap in SAMA at age nine - excluding monozygotic twins from the sample

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 3 OS subsample	(5) Model 4
Female	-0.445***	-0.384***	-0.378***	-0.449***	-0.447***
	(0.040)	(0.037)	(0.037)	(0.051)	(0.052)
Math level, age 9		0.374***	0.336***	0.319***	0.395***
		(0.019)	(0.021)	(0.030)	(0.036)
Verbal abilities, age 9			0.028	-0.030	0.047
			(0.022)	(0.030)	(0.044)
Non-verbal abilities, age 9			0.071***	0.110***	0.140***
			(0.024)	(0.034)	(0.041)
Constant	0.249***	0.170***	0.022	0.251*	0.202
	(0.040)	(0.037)	(0.109)	(0.151)	(0.317)
Observations	2,436	2,436	2,436	1,186	2,436
R-squared	0.050	0.183	0.190	0.195	0.211
Twin FE	No	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	Yes	No

Source: TEDS (Rimfeld et al., 2019). Sample of dizygotic twins only. Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: elder twin, heavier twin, and birth weight.

Table B5: The gender gap in SAMA at age 12, OLS models

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5
Female	-0.394***	-0.337***	-0.299***	-0.340***	-0.285***
	(0.038)	(0.034)	(0.033)	(0.075)	(0.082)
Math level, age 12		0.459***	0.296***	0.228***	0.184***
		(0.023)	(0.023)	(0.048)	(0.054)
Math test scores, age 12			0.022***	0.014***	0.012***
			(0.002)	(0.004)	(0.004)
Verbal abilities, age 12			-0.008	-0.026	-0.068
			(0.020)	(0.047)	(0.057)
Non-verbal abilities, age 12			-0.015	0.044	0.047
			(0.020)	(0.049)	(0.050)
Math level, age 9				0.216***	0.191***
				(0.045)	(0.055)
Verbal abilities, age 9				0.037	0.018
				(0.041)	(0.045)
Non-verbal abilities, age 9				0.059	0.100*
				(0.049)	(0.056)
Math level, age 7					0.125**
					(0.059)
Verbal abilities, age 7					0.046
					(0.053)
Non-verbal abilities, age 7					-0.020
					(0.046)
Elder twin			0.029	0.006	0.030
			(0.027)	(0.062)	(0.067)
Heavier twin at birth			0.024	0.128*	0.131*
			(0.030)	(0.067)	(0.073)
Birth weight, grams			0.000	-0.000	-0.000
			(0.000)	(0.000)	(0.000)
Constant	0.272***	-0.004	-1.649***	-0.871***	-0.792**
	(0.058)	(0.054)	(0.142)	(0.322)	(0.335)
Observations	3,196	3,196	3,196	570	460
R-squared	0.038	0.205	0.263	0.343	0.361
Twin FE	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B6: The gender gap in SAMA at age 12 - FE models

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5
Female	-0.482***	-0.448***	-0.413***	-0.536***	-0.464***
	(0.065)	(0.060)	(0.060)	(0.134)	(0.138)
Math level, age 12		0.547***	0.420***	0.284***	0.148*
		(0.043)	(0.042)	(0.076)	(0.084)
Math test scores, age 12			0.019***	0.018**	0.012
			(0.003)	(0.007)	(0.008)
Verbal abilities, age 12			0.069**	0.016	-0.065
			(0.034)	(0.094)	(0.100)
Non-verbal abilities, age 12			-0.006	0.093	0.087
			(0.031)	(0.067)	(0.069)
Math level, age 9				0.247***	0.226**
				(0.079)	(0.090)
Verbal abilities, age 9				0.016	-0.009
				(0.083)	(0.098)
Non-verbal abilities, age 9				-0.070	-0.048
				(0.089)	(0.108)
Math level, age 7					0.342***
					(0.113)
Verbal abilities, age 7					0.192**
					(0.078)
Non-verbal abilities, age 7					-0.057
					(0.070)
Elder twin			0.030	-0.031	0.010
			(0.028)	(0.065)	(0.071)
Heavier twin at birth			0.045	0.179*	0.083
			(0.047)	(0.105)	(0.115)
Birth weight, grams			-0.000	-0.000	-0.000
			(0.000)	(0.000)	(0.000)
Constant	0.277***	0.225***	-1.012***	-0.307	-0.527
	(0.037)	(0.035)	(0.338)	(0.702)	(0.712)
Observations	3,196	3,196	3,196	570	460
R-squared	0.043	0.178	0.224	0.301	0.321
Twin FE	Yes	Yes	Yes	Yes	Yes
Cohort FE	No	No	No	No	No

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B7: SAMA as a categorical variable, age nine (Multinomial logit model)

VARIABLES	(1) SAMA=1	(2) SAMA=2	(3) SAMA=3	(4) SAMA=4	(5) SAMA=5
Female	0.251	0.107		-0.398***	-1.021***
	(0.180)	(0.121)		(0.085)	(0.109)
Math level, age 9	-0.723***	-0.283***		0.353***	0.757***
	(0.102)	(0.068)		(0.049)	(0.066)
Verbal abilities, age 9	-0.096	-0.116*		0.058	0.070
	(0.092)	(0.066)		(0.047)	(0.063)
Non-verbal abilities, age 9	-0.162	-0.079		0.038	0.096
	(0.099)	(0.065)		(0.048)	(0.066)
Constant	-2.274***	-1.161***		0.109	-0.738**
	(0.510)	(0.313)		(0.227)	(0.287)
Observations	3,877	3,877	3,877	3,877	3,877

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The estimated multinomial logit model is the following: $f(k,i) = beta_k * x_i$, where $beta_k$ is a set of regression coefficients associated with (integer) SAMA values k, $k = 1, 2, .., 5$, and x_i is the same set of explanatory variables associated with observation i as before. $SAMA = 3$ is the baseline category. Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table B8: The role of co-twin (CT) SAMA, age 12 sample

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.279***	-0.280***	-0.345***	-0.316***		
	(0.033)	(0.033)	(0.039)	(0.036)		
Has a male twin (MT)			-0.156***	-0.140***	-0.116**	-0.164***
			(0.040)	(0.037)	(0.051)	(0.052)
SAMA of CT, age 12, std	0.132***	0.117***	0.139***	0.029	0.023	0.181***
	(0.024)	(0.032)	(0.029)	(0.044)	(0.043)	(0.036)
MT*SAMA of CT			0.015	0.163***	0.169***	-0.098
			(0.038)	(0.059)	(0.060)	(0.060)
Female*SAMA of CT		0.027		0.150***		
		(0.037)		(0.054)		
Female*MT*SAMA of CT				-0.266***		
				(0.103)		
Constant	-1.499***	-1.500***	-1.403***	-1.414***	-1.449***	-1.691***
	(0.141)	(0.141)	(0.143)	(0.141)	(0.213)	(0.183)
Observations	3,012	3,012	3,012	3,012	1,275	1,737
R-squared	0.279	0.279	0.283	0.287	0.279	0.252
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Further control variables: mathematics level and mathematics test scores at age 12, verbal and non-verbal cognitive skills at age 12, elder twin, heavier twin, and birth weight.

Table B9: The role of co-twin (CT) SAMA - using age seven math levels as an IV for age nine math levels

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.305***	-0.305***	-0.373***	-0.331***		
	(0.035)	(0.036)	(0.039)	(0.038)		
Math level, age 9	0.431***	0.431***	0.425***	0.419***	0.445***	0.383***
	(0.043)	(0.043)	(0.042)	(0.042)	(0.060)	(0.057)
Has a male twin (MT)			-0.177***	-0.149***	-0.177***	-0.120**
			(0.041)	(0.037)	(0.053)	(0.053)
SAMA of CT, age 9, std	0.129***	0.100***	0.167***	0.015	0.019	0.231***
	(0.026)	(0.034)	(0.033)	(0.043)	(0.043)	(0.042)
MT*SAMA of CT			-0.050	0.147**	0.140**	-0.226***
			(0.044)	(0.062)	(0.062)	(0.065)
Female*SAMA of CT		0.055		0.212***		
		(0.042)		(0.058)		
Female*MT*SAMA of CT				-0.362***		
				(0.107)		
Constant	-0.076	-0.073	0.037	0.011	0.029	-0.314**
	(0.096)	(0.097)	(0.099)	(0.098)	(0.138)	(0.125)
Observations	2,824	2,824	2,824	2,824	1,296	1,528
R-squared	0.182	0.183	0.190	0.198	0.184	0.162
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Further control variables: mathematics level and mathematics test scores at age 12, verbal and non-verbal cognitive skills at age 12, elder twin, heavier twin, and birth weight. 2SLS estimates using age seven math levels as instrumental variables for age nine math levels.

Table B10: The role of co-twin (CT) SAMA - using the ORIV approach of Gillen et al. (2019)

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.295***	-0.295***	-0.351***	-0.312***		
	(0.035)	(0.035)	(0.039)	(0.037)		
Math levels	0.529***	0.529***	0.524***	0.518***	0.502***	0.536***
	(0.040)	(0.040)	(0.040)	(0.040)	(0.052)	(0.060)
Has a male twin (MT)			-0.147***	-0.122***	-0.130**	-0.115**
			(0.040)	(0.037)	(0.052)	(0.053)
SAMA of CT, age 9, std	0.114***	0.086***	0.152***	0.007	0.010	0.206***
	(0.026)	(0.033)	(0.032)	(0.043)	(0.043)	(0.042)
MT*SAMA of CT			-0.056	0.134**	0.134**	-0.217***
			(0.043)	(0.061)	(0.061)	(0.064)
Female*SAMA of CT		0.053		0.201***		
		(0.041)		(0.058)		
Female*MT*SAMA of CT				-0.349***		
				(0.104)		
Constant	0.008	0.010	0.100	0.073	0.059	-0.220*
	(0.097)	(0.097)	(0.099)	(0.097)	(0.134)	(0.127)
Observations	5,876	5,876	5,876	5,876	2,696	3,180
R-squared	0.113	0.114	0.121	0.130	0.129	0.079
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Further control variables: mathematics level and mathematics test scores at age 12, verbal and non-verbal cognitive skills at age 12, elder twin, heavier twin, and birth weight. 2SLS estimates using the ORIV approach of Gillen et al. (2019).

Table B11: The role of co-twin (CT) SAMA - excluding monozygotic twins from the sample

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.390***	-0.392***	-0.393***	-0.382***		
	(0.038)	(0.038)	(0.038)	(0.039)		
Has a male twin (MT)			-0.088**	-0.085**	-0.052	-0.117**
			(0.039)	(0.038)	(0.054)	(0.055)
SAMA of CT, age 9, std	0.053**	0.032	0.082**	0.035	0.030	0.128**
	(0.027)	(0.032)	(0.032)	(0.036)	(0.036)	(0.051)
MT*SAMA of CT			-0.042	0.016	0.011	-0.093
			(0.041)	(0.067)	(0.068)	(0.067)
Female*SAMA of CT		0.041		0.088		
		(0.039)		(0.061)		
Female*MT*SAMA of CT				-0.110		
				(0.107)		
Constant	0.029	0.032	0.068	0.066	0.029	-0.286*
	(0.109)	(0.109)	(0.110)	(0.110)	(0.148)	(0.150)
Observations	2,328	2,328	2,328	2,328	1,119	1,209
R-squared	0.195	0.195	0.197	0.198	0.178	0.141
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Sample of dizygotic twins only. Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. CT refers to co-twins.

Table B12: The role of stereotypically gender-biased parental assessments in the gender gap in SAMA - alternative model

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.324***	-0.303***	0.033	-0.447***	-0.442***	-0.052
	(0.032)	(0.032)	(0.037)	(0.051)	(0.052)	(0.061)
Stereotipically assessed person, 5 cat		-0.127***	0.577***		-0.048	0.586***
		(0.034)	(0.045)		(0.051)	(0.080)
Female*stereotipically assessed, 5 cat			-1.144***			-1.137***
			(0.063)			(0.111)
Constant	-0.064	-0.035	-0.114	0.033	0.040	-0.069
	(0.089)	(0.089)	(0.085)	(0.236)	(0.237)	(0.227)
Observations	3,877	3,877	3,877	3,877	3,877	3,877
R-squared	0.174	0.177	0.240	0.164	0.164	0.216
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics levels at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. The alternative measure of parental stereotypical assessment was created the following way. First, a multinomial logit model of the form $f(k,i) = beta_k * x_i$ is estimated, where $beta_k$ is a set of regression coefficients associated with k categories of parental assessments, $k = 1, 2, 3, 4, 5$, and x_i is the set of three explanatory variables: mathematics levels and verbal and non-verbal cognitive skills at age nine. Then, predicted categories of parental assessment are fitted by the model and they are compared to the observed parental assessment of individuals. An individual is over(under)estimated if their observed parental assessment category is higher(lower) than their predicted category.

Table B13: The role of stereotypically gender-biased parental assessments in the gender gap in SAMA - using age seven math levels as an IV for age nine math levels

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.297***	-0.297***	-0.001	-0.457***	-0.458***	-0.112
	(0.037)	(0.037)	(0.043)	(0.063)	(0.062)	(0.074)
Math level, age 9	0.475***	0.475***	0.461***	0.717***	0.719***	0.675***
	(0.043)	(0.043)	(0.041)	(0.110)	(0.110)	(0.103)
Stereotipically assessed person		-0.005	0.619***		-0.034	0.561***
		(0.041)	(0.051)		(0.065)	(0.092)
Female*stereotipically assessed			-1.157***			-1.143***
			(0.082)			(0.144)
Constant	-0.089	-0.088	-0.196**	0.108	0.109	-0.043
	(0.103)	(0.104)	(0.099)	(0.280)	(0.280)	(0.267)
Observations	2,942	2,942	2,942	2,942	2,942	2,942
R-squared	0.160	0.160	0.225			
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics levels at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. 2SLS estimates using age seven math levels as instrumental variables for age nine math levels. The measure of parental stereotypical assessment was created the following way. First, a multinomial logit model of the form $f(k,i) = beta_k * x_i$ is estimated, where $beta_k$ is a set of regression coefficients associated with the k terciles of parental assessments, $k = 1, 2, 3$, and x_i is the set of three explanatory variables: mathematics levels and verbal and non-verbal cognitive skills at age nine. Then, predicted categories of parental assessment are fitted by the model and they are compared to the observed parental assessment of individuals. An individual is over(under)estimated if their observed parental assessment category is higher(lower) than their predicted category.

Table B14: The role of stereotypically gender-biased parental assessments in the gender gap in SAMA - using the ORIV approach of Gillen et al. (2019)

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.291***	-0.291***	-0.036	-0.604***	-0.600***	-0.346***
	(0.036)	(0.036)	(0.043)	(0.075)	(0.075)	(0.090)
Math levels	0.568***	0.569***	0.598***	-0.601***	-0.594***	-0.601***
	(0.039)	(0.039)	(0.039)	(0.097)	(0.096)	(0.097)
Stereotipically assessed person		-0.010	0.515***		0.131*	0.569***
		(0.038)	(0.050)		(0.073)	(0.113)
Female*stereotipically assessed			-0.985***			-0.845***
			(0.075)			(0.165)
Constant	0.006	0.009	-0.069	0.208	0.198	0.083
	(0.102)	(0.103)	(0.100)	(0.338)	(0.338)	(0.333)
Observations	6,118	6,118	6,118	6,118	6,118	6,118
R-squared	0.085	0.085	0.114			
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics levels at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. 2SLS estimates using age seven math levels as instrumental variables for age nine math levels. 2SLS estimates using the ORIV approach of Gillen et al. (2019). The measure of parental stereotypical assessment was created the following way. First, a multinomial logit model of the form $f(k,i) = beta_k * x_i$ is estimated, where $beta_k$ is a set of regression coefficients associated with the k terciles of parental assessments, $k = 1, 2, 3$, and x_i is the set of three explanatory variables: mathematics levels and verbal and non-verbal cognitive skills at age nine. Then, predicted categories of parental assessment are fitted by the model and they are compared to the observed parental assessment of individuals. An individual is over(under)estimated if their observed parental assessment category is higher(lower) than their predicted category.

Table B15: The role of stereotypically gender-biased parental assessments in the gender gap in SAMA - excluding monozygotic twins from the sample

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.378***	-0.378***	-0.047	-0.447***	-0.446***	-0.060
	(0.037)	(0.038)	(0.045)	(0.052)	(0.052)	(0.065)
Stereotypically assessed person		-0.000	0.572***		0.035	0.638***
		(0.040)	(0.051)		(0.059)	(0.084)
Female*stereotypically assessed			-1.165***			-1.236***
			(0.082)			(0.134)
Constant	0.022	0.022	-0.056	0.202	0.191	-0.000
	(0.109)	(0.110)	(0.106)	(0.317)	(0.318)	(0.301)
Observations	2,436	2,436	2,436	2,436	2,436	2,436
R-squared	0.190	0.190	0.253	0.211	0.211	0.271
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Sample of dizygotic twins only.Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics levels at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. The measure of parental stereotypical assessment was created the following way. First, a multinomial logit model of the form $f(k,i) = beta_k * x_i$ is estimated, where $beta_k$ is a set of regression coefficients associated with the k terciles of parental assessments, $k = 1, 2, 3$, and x_i is the set of three explanatory variables: mathematics levels and verbal and non-verbal cognitive skills at age nine. Then, predicted categories of parental assessment are fitted by the model and they are compared to the observed parental assessment of individuals. An individual is over(under)estimated if their observed parental assessment category is higher(lower) than their predicted category.

C Deviations from pre-registration protocol

As mentioned in the acknowledgments to the paper, this study was pre-registered in the OSF Registries (https://osf.io/chv5g). This is a pre-requisite of obtaining TEDS data and must be completed before data access is granted.

The analysis in this paper deviates from the pre-registration in three key ways. The first deviation is that we restricted the focus of the study from ambition, risk-taking, and overconfidence to overconfidence. This was due to the volume of results and the desire to keep the paper simple.

The second is that we decided not to construct the composite measure of overconfidence and instead focus just on mathematics self-assessment controlling for actual mathematics ability. In the psychological literature, there are two main ways of capturing overconfidence. One is to construct an overconfidence measure (either a residual score or a difference measure) using measures of self-assessment and actual ability. The other is to compare self-assessments conditional on actual ability. We had initially wanted to construct a similar overconfidence measure to Adamecz-Völgyi and Shure (2022), but TEDS did not have the range of measures to do this. The overconfidence measure we could have constructed would have been based only on English and mathematics self-assessments and on English and mathematics national curriculum levels (actual performance). Given that the gender and ability gaps work in opposite directions with these two measures, we would have ended up with an overconfidence measure that had zero gender differences. We decided instead to follow the second approach and look at conditional mathematics self-assessments since this was the category with the largest gender gap in favor of boys and the domain most important for future labor market success.

The third deviation is that we did not undertake the Kitagawa-Blinder-Oaxaca decomposition, but instead focused on linear regressions to assess the gender gap. Both of these methods were outlined in the protocol, but in the interest of brevity, we focus on the linear regression results.

Online Appendix to "Peers, parents, and self-perceptions: The gender gap in mathematics self-assessment"

O1 Attrition and non-response

As detailed in the paper, our main analytical sample contains 3,877 observations out of the total initial sample of Cohort (1) and (2) of TEDS (15,216 observations). In this section of the Online Appendix, we provide robustness checks to show that selection to this subsample of TEDS is not likely to bias our results.

Table O1 compares those in our analytical sample to those who participated in the first wave, but either dropped out by age nine or they did not provide all data we needed. Those in our analytical sample come from slightly better social backgrounds: their parents are more likely to have qualifications, work in better jobs, and their fathers were more likely to live with the family right after when they were born.

Table O2 looks at selection to the analytic sample using a linear probability and a probit model. SES is positively correlated with the probability of being in the sample, while missing data (i.e., non-response to some questions already in the first wave) is negatively correlated with it. Interestingly, those with younger siblings are also less likely to be in the sample.

Table O1: The differences between those in the analytical sample and those who dropped out

	Mean, dropouts	Mean, analytical sample	Diff	p-value
No father in family	0.1	0.06	0.05	0
Qual needed for job, mother	0.17	0.25	-0.09	0
Qual needed for job, father	0.39	0.47	-0.09	0
Family SES score	-0.13	0.17	-0.3	0
Family SES missing	0.1	0.04	0.06	0
Age of mother	30.28	31.38	-1.1	0
Age of mother missing	0.02	0.01	0.01	0
Cohort: 2	1.59	1.62	-0.03	0
Has younger siblings	0.03	0.02	0.02	0
Has older siblings	0.54	0.5	0.04	0
Father no qual	0.14	0.09	0.05	0
Mother no qual	0.12	0.06	0.07	0
Mother's qual missing	0	0	0	0
Emp of mother: manager	0.06	0.09	-0.03	0
Emp of mother: employee	0.28	0.33	-0.05	0
Emp of mother: SE with emps	0.02	0.01	0	0.237
Emp of mother: SE	0.03	0.04	-0.01	0.003
Emp of father: manager	0.22	0.29	-0.07	0
Emp of father: employee	0.36	0.38	-0.02	0.003
Emp of father: SE with emps	0.07	0.08	0	0.443
Emp of father: SE	0.1	0.1	0	0.265
Emp of father: Foreman	0	0	0	0.072
Ethnicity: White	0.9	0.95	-0.05	0
Ethnicity: Missing	0	0	0	0.026

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p«0.01, ** p«0.05, * p«0.1 Sample of those in the first wave. No. of observations: 3,877 in the analytical sample, 11,339 dropped out or did not provide all data at age nine. Note that there are also differences between the two groups in their ACORN codes, but those are not reported. ACORN captures geodemographic neighborhood characteristics.

Table O2: Selection to the analytical sample

PART 1	(1) Linear probability model	(2) Probit
No father in family	0.031	0.127
	(0.024)	(0.094)
Mother needs qualification for job	0.052***	0.145***
	(0.016)	(0.048)
Father needs qualification for job	0.007	0.023
	(0.011)	(0.035)
SES score	0.035***	0.107***
	(0.008)	(0.026)
SES score missing	-0.078***	-0.318***
	(0.024)	(0.100)
Mother's age	0.003***	0.013***
	(0.001)	(0.004)
Mother's age missing	-0.053	-0.258
	(0.034)	(0.157)
School cohort	0.022**	0.071**
	(0.010)	(0.032)
Has younger sibling	-0.064***	-0.273**
	(0.024)	(0.108)
Has older sibling	-0.011	-0.043
	(0.010)	(0.034)
Father has no qualification	-0.009	-0.041
	(0.016)	(0.057)
Mother has no qualification	-0.060***	-0.269***
	(0.015)	(0.064)
Data on mother's qualification is missing	-0.051	-0.419
	(0.058)	(0.514)
Mother's work: manager	0.003	0.010
	(0.024)	(0.070)
Mother's work: employee	0.007	0.023
	(0.013)	(0.041)
Mother's work: SE with employees	-0.072*	-0.243*
	(0.038)	(0.130)
Mother's work: Foreman	-0.059	-0.188
	(0.089)	(0.292)
Mother's work: SE without employees	-0.284***	
	(0.055)	
Father's work: manager	0.023	0.082
	(0.016)	(0.052)
Father's work: employee	0.033**	0.112**
	(0.014)	(0.046)

PART 2	(1) Linear probability model	(2) Probit
Father's work: SE with employees	0.006	0.031
	(0.021)	(0.070)
Father's work: Foreman	0.003	0.017
	(0.028)	(0.098)
Father's work: SE without employees	0.014	0.067
	(0.133)	(0.632)
Ethnicity: White	0.086***	0.333***
	(0.017)	(0.070)
Ethnicity: missing	0.033	0.139
	(0.076)	(0.330)
Constant	-0.047	-1.783***
	(0.044)	(0.161)
Observations	15,216	15,162
R-squared	0.055	

Source: TEDS (Rimfeld et al., 2019). Sample of those in the first wave. Robust standard errors clustered within twin pairs in parentheses. *** p«0.01, ** p«0.05, * p«0.1 We also control ACORN codes in both models.

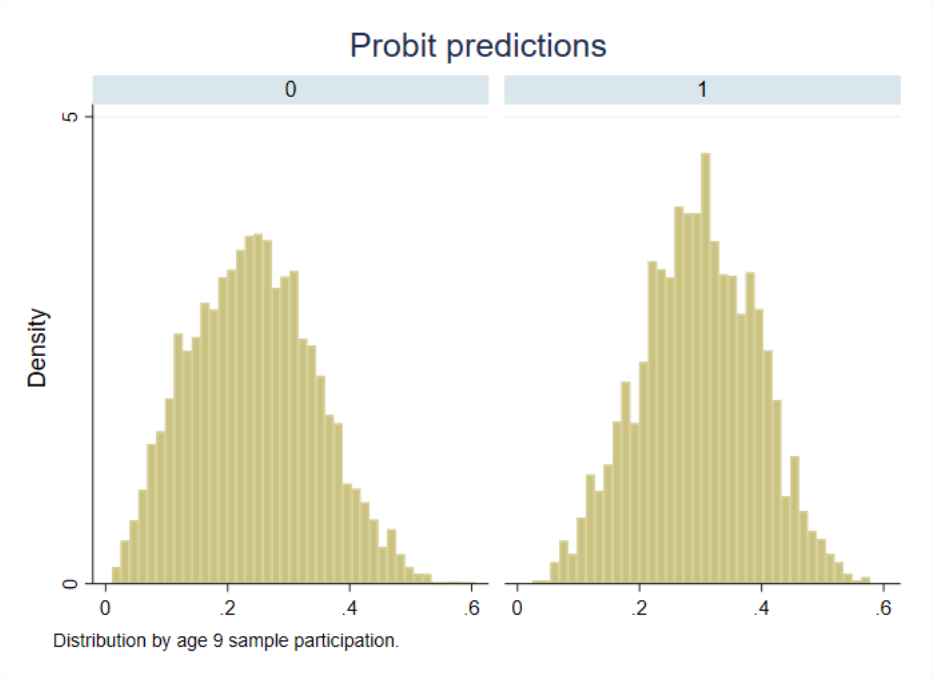
First, we use the estimated probabilities of being in the analytical sample from the probit model of Table O2 to create inverse probability weights (IPW). This method ensures that those with higher probability (i.e., those from higher SES backgrounds) get lower weights, so we compensate for them being less likely to drop out. Using these weights, we re-estimate our three most interesting results: our main results as in Table 1 in the main text, the role of parental stereotypical evaluations as in Table 5 in the main text, and the role of male co-twin SAMA as in Table 3 in the main text. These results are reported along with the results of the other two re-weighting methods in Tables O3, O4 and O5 (Block A).

Second, we re-estimate the selection model using a non-parametric machine learning algorithm, random forest. This works by constructing a series of decision trees and predicting the outcome from each series as the modes of predictions (Breiman, 2001). The method offers several advantages. First, as it randomly splits the sample along the explanatory variables, it explicitly models potential non-linear relationships. Thus, if non-linearities are important, it offers better predictions than a probit (where all parameters are linear). Indeed, comparing the (in-sample) predictive power of the probit and the random forest models, the random forest provides almost 50% higher AUC (a measure of predictive power, Kuhn and Johnson (2013)) than the probit (0.99 vs 0.65).[10]

The second advantage of a random forest classification algorithm, besides giving a better prediction, is that it ranks the predictors in terms of their importance (Grömping, 2009), helping us to understand more how selection works. Figure O2 shows the estimated importance measures.

[10]The value of AUC is between 0 and 1, and flipping a coin would produce an AUC of 0.5. As a rule of thumb, the predictive power of a model is considered good if AUC>0.8 and great if AUC>0.9.

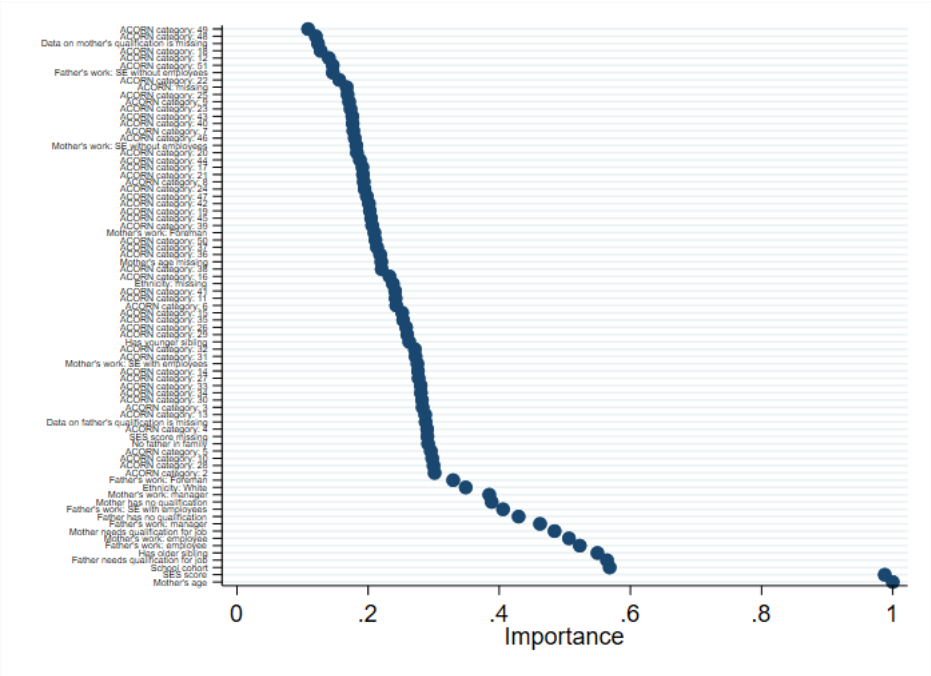
Figure O1: The predicted probability of being in the analytical sample (probit model)



Source: TEDS (Rimfeld et al., 2019). No. of obs: 15,162. Robust standard errors clustered within twin pairs in parentheses. *** p«0.01, ** p«0.05, * p«0.1 The left panel shows the estimated probabilities for those who dropped out, while the right panel for those who are in the analytical sample.

Interestingly, the mother's age and family SES scores are the most important predictors of being selected to the analytical sample, followed by the measures of parents' employment measures.

Figure O2: The importance of predictors in predicting selection to the analytical sample in a random forest model

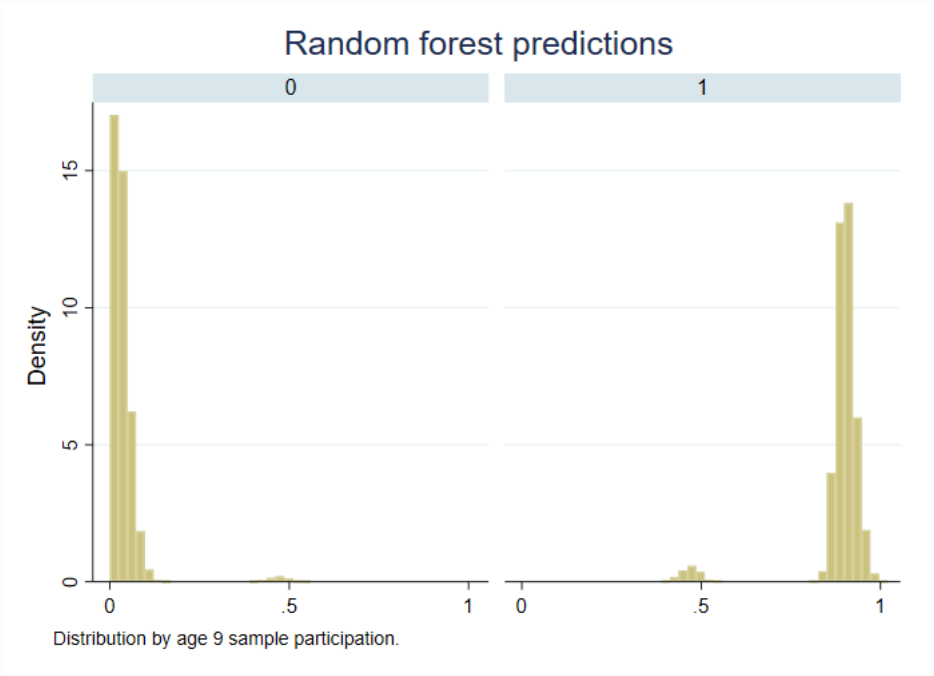


Source: TEDS (Rimfeld et al., 2019). No. of obs: 15,162. Importance shows the relative predictive power of each explanatory variable compared to the predictive power of the most important variable (importance=1). We measure the relative importance of explanatory variables by the Mean Decrease in Gini measure, which captures how well the variable decreases the heterogeneity of subgroups by splitting the sample on a given variable averaged across all decision trees (Friedman et al., 2009).

Third, while the probit model assumes a normal distribution for the predicted probabilities, the random forest does not. Thus, the predictions themselves are quite different (compare Figure O1 and Figure O3). This is useful for us because it is reassuring that our results do not change with either type of re-weighting. Similarly to the probit model, we take the inverse of these probabilities to create IPW's. Our main results re-estimated using these weights are reported in Tables O3, O4 and O5 (Block B).

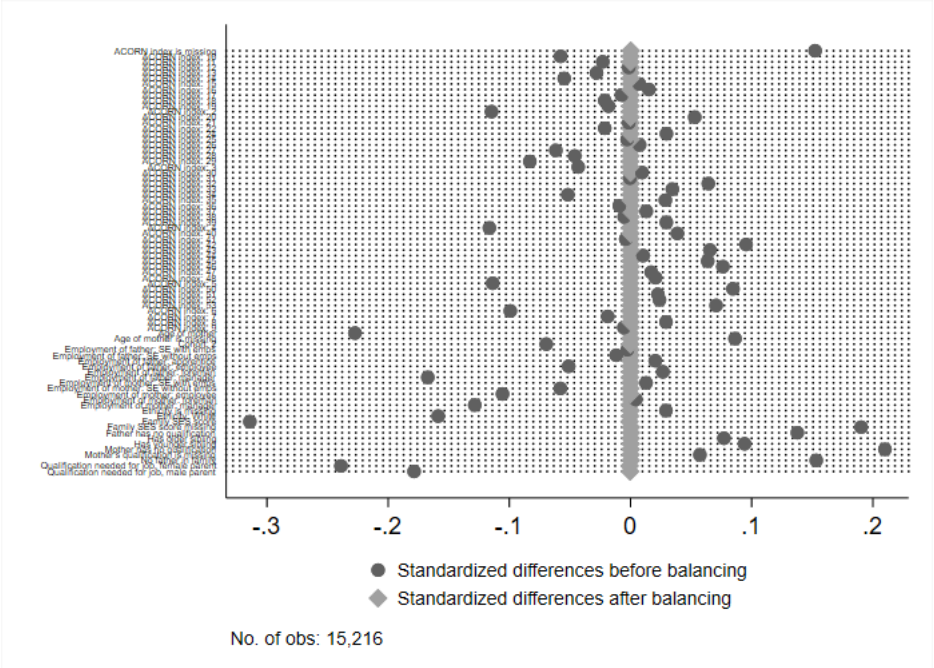
Lastly, we use entropy balancing (Hainmueller, 2012) to re-weight the analytical sample in a way that the observed characteristics of the sample members follow the distribution of characteristics among those who dropped out. Figure O4 shows the balance of characteristics before and after applying the entropy balanced weights. Using these weights eliminates statistical differences between those in the analytical sample and those who were excluded. Our main results re-estimated using these weights are reported in Tables O3, O4 and O5 (Block C).

Figure O3: The predicted probability of being in the analytical sample (random forest model)



Source: TEDS (Rimfeld et al., 2019). Number of observations: 15,162. The left panel shows the estimated probabilities for those who dropped out, while the right panel for those who are in the analytical sample.

Figure O4: The balance of the analytical sample compared to those who dropped out before and after using entropy-balanced weights)



Source: TEDS (Rimfeld et al., 2019). Number of observations: 15,162. Note that after applying the entropy-balanced weights, the standardized differences fall very close to zero; hence the individual points collapse to a vertical line at x=0.

Table O3: The gender gap in mathematics self-assessment (SAMA) - weighted results

	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
Block A: Weighted using probit IPW				
Female	-0.355***	-0.309***	-0.306***	-0.413***
	(0.038)	(0.035)	(0.035)	(0.058)
Block B: Weighted using random forest IPW				
Female	-0.368***	-0.320***	-0.317***	-0.446***
	(0.035)	(0.032)	(0.032)	(0.052)
Block C: Weighted using entropy-balanced weights				
Female	-0.342***	-0.296***	-0.294***	-0.393***
	(0.042)	(0.040)	(0.040)	(0.064)
Observations	3,877	3,877	3,877	3,877
Control variables				
Math levels		Yes	Yes	Yes
Verbal and non-verbal abilities			Yes	Yes
Birth characteristics			Yes	Yes
Twin FE				Yes
Cohort FE	Yes	Yes	Yes	

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table O4: The role of stereotypically gender-biased parental assessments in the gender gap in SAMA - weighted results

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Block A: Weighted using probit IPW						
Female	-0.306***	-0.306***	-0.015	-0.413***	-0.413***	-0.134**
	(0.035)	(0.035)	(0.042)	(0.058)	(0.058)	(0.068)
Stereotipically assessed person		0.009	0.559***		0.007	0.446***
		(0.038)	(0.049)		(0.059)	(0.086)
Female*stereotipically assessed			-1.085***			-0.889***
			(0.073)			(0.129)
Block B: Weighted using random forest IPW						
Female	-0.316***	-0.316***	-0.019	-0.446***	-0.446***	-0.146**
	(0.032)	(0.032)	(0.037)	(0.052)	(0.052)	(0.061)
Stereotipically assessed person		0.012	0.592***		0.025	0.508***
		(0.034)	(0.043)		(0.051)	(0.076)
Female*stereotipically assessed			-1.108***			-0.956***
			(0.066)			(0.116)
Block C: Weighted using entropy-balanced weights						
Female	-0.294***	-0.293***	-0.003	-0.393***	-0.393***	-0.121
	(0.040)	(0.040)	(0.046)	(0.064)	(0.065)	(0.076)
Stereotipically assessed person		0.016	0.551***		-0.005	0.415***
		(0.043)	(0.057)		(0.064)	(0.092)
Female*stereotipically assessed			-1.082***			-0.875***
			(0.081)			(0.141)
Observations	3,877	3,877	3,877	3,877	3,877	3,877
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table O5: The role of co-twin (CT) SAMA - weighted results

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Block A: probit IPW						
Female	-0.306***	-0.306***	-0.360***	-0.318***		
	(0.034)	(0.034)	(0.039)	(0.037)		
Has a male twin (MT)			-0.138***	-0.113***	-0.114**	-0.104**
			(0.041)	(0.037)	(0.050)	(0.053)
SAMA of CT, age 9, std	0.146***	0.116***	0.176***	0.017	0.017	0.242***
	(0.025)	(0.031)	(0.032)	(0.038)	(0.038)	(0.042)
MT*SAMA of CT			-0.043	0.168***	0.161***	-0.225***
			(0.043)	(0.057)	(0.057)	(0.063)
Female*SAMA of CT		0.056		0.221***		
		(0.041)		(0.055)		
Female*MT*SAMA of CT				-0.390***		
				(0.101)		
Block B: random forest IPW						
Female	-0.324***	-0.324***	-0.382***	-0.341***		
	(0.031)	(0.031)	(0.035)	(0.033)		
Has a male twin (MT)			-0.151***	-0.127***	-0.126***	-0.122***
			(0.036)	(0.033)	(0.045)	(0.047)
SAMA of CT, age 9, std	0.153***	0.127***	0.182***	0.032	0.031	0.246***
	(0.023)	(0.029)	(0.028)	(0.037)	(0.036)	(0.036)
MT*SAMA of CT			-0.034	0.165***	0.161***	-0.208***
			(0.039)	(0.054)	(0.055)	(0.057)
Female*SAMA of CT		0.048		0.209***		
		(0.037)		(0.050)		
Female*MT*SAMA of CT				-0.371***		
				(0.094)		
Block C: entropy balanced weights						
Female	-0.295***	-0.294***	-0.346***	-0.304***		
	(0.038)	(0.038)	(0.043)	(0.041)		
Has a male twin (MT)			-0.130***	-0.104**	-0.119**	-0.082
			(0.045)	(0.041)	(0.058)	(0.057)
SAMA of CT, age 9, std	0.140***	0.111***	0.168***	0.005	0.008	0.235***
	(0.027)	(0.034)	(0.035)	(0.041)	(0.041)	(0.046)
MT*SAMA of CT			-0.041	0.175***	0.166***	-0.235***
			(0.047)	(0.062)	(0.063)	(0.069)
Female*SAMA of CT		0.053		0.226***		
		(0.045)		(0.060)		
Female*MT*SAMA of CT				-0.404***		
				(0.111)		
Observations	3,722	3,722	3,722	3,722	1,707	2,015

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, birth weight and cohort FE.

O2 Supporting information

Table O6: The gender gap in SAMA, fully interacted main model, age nine

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
Female	-0.376***	-0.325***	-0.174	-0.188
	(0.034)	(0.032)	(0.160)	(0.262)
Math level, age 9		0.387***	0.326***	0.356***
		(0.022)	(0.018)	(0.032)
Female*math level		-0.030		
		(0.033)		
Verbal abilities, age 9			0.074***	0.139***
			(0.026)	(0.049)
Female*verbal abilities			-0.038	-0.101*
			(0.036)	(0.058)
Non-verbal abilities, age 9			0.069**	0.123***
			(0.027)	(0.044)
Female*nonverbal abilities			-0.016	0.010
			(0.037)	(0.056)
Elder twin			0.050	-0.003
			(0.038)	(0.042)
Female*elder twin			-0.024	0.065
			(0.054)	(0.062)
Heavier twin at birth			0.024	0.028
			(0.042)	(0.060)
Female*heavier twin			0.033	0.010
			(0.058)	(0.073)
Birthweight, gramms			0.000*	0.000
			(0.000)	(0.000)
Female*birthweight			-0.000	-0.000
			(0.000)	(0.000)
Constant	0.182***	0.116***	-0.140	-0.140
	(0.035)	(0.032)	(0.116)	(0.287)
Observations	3,877	3,877	3,877	3,877
R-squared	0.036	0.165	0.175	0.167
Twin FE	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	No

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table O7: The gender gap in mathematics levels, age nine

VARIABLES	(1) Model 1	(2) Model 3	(3) Model 4	(4) Model 5	(5) Model 6	(6) Model 7
Female	-0.133***	-0.106***	-0.174***	-0.129***	-0.160***	-0.128***
	(0.034)	(0.030)	(0.041)	(0.028)	(0.045)	(0.028)
Has a male twin				-0.059**	-0.092**	-0.059**
				(0.028)	(0.047)	(0.028)
Has brother						-0.041
						(0.036)
Has sister						-0.022
						(0.036)
Verbal abilities, age 9		0.214***	0.202***	0.214***	0.213***	0.212***
		(0.017)	(0.026)	(0.017)	(0.017)	(0.017)
Non-verbal abilities, age 9		0.302***	0.243***	0.303***	0.303***	0.302***
		(0.017)	(0.026)	(0.017)	(0.017)	(0.017)
Elder twin		0.015	0.013	0.015	0.014	0.015
		(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Heavier twin at birth		0.053**	0.091***	0.047**	0.048**	0.045**
		(0.023)	(0.033)	(0.023)	(0.023)	(0.023)
Birthweight, gramms		0.000***	-0.000	0.000***	0.000***	0.000***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female*male twin					0.064	
					(0.069)	
Constant	0.148***	-0.219***	0.130	-0.181**	-0.153*	-0.179**
	(0.036)	(0.081)	(0.191)	(0.082)	(0.089)	(0.082)
Observations	4,309	4,309	4,309	4,309	4,309	4,309
R-squared	0.005	0.202	0.129	0.203	0.203	0.203
Twin FE	No	No	Yes	No	No	No
Cohort FE	Yes	Yes	No	Yes	Yes	Yes

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table O8: The gender gap in self-assessed English abilities, age nine

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
Female	0.224***	0.132***	0.141***	0.231***
	(0.035)	(0.032)	(0.032)	(0.053)
English level, age 9		0.356***	0.333***	0.382***
		(0.018)	(0.020)	(0.038)
Verbal abilities, age 9			0.095***	0.119***
			(0.019)	(0.038)
Non-verbal abilities, age 9			-0.032	0.007
			(0.019)	(0.037)
Constant	-0.113***	-0.106***	-0.054	-0.203
	(0.034)	(0.031)	(0.086)	(0.229)
Observations	3,877	3,877	3,877	3,877
R-squared	0.012	0.128	0.135	0.107
Twin FE	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	No

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table O9: The gender gap in parental assessment of their children's mathematics abilities

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 3 OS subsample	(5) Model 4
Female	-0.289***	-0.212***	-0.209***	-0.400***	-0.417***
	(0.037)	(0.030)	(0.029)	(0.043)	(0.042)
Math level, age 9		0.591***	0.517***	0.490***	0.499***
		(0.016)	(0.017)	(0.029)	(0.028)
Verbal abilities, age 9			0.082***	0.047	0.094***
			(0.017)	(0.031)	(0.026)
Non-verbal abilities, age 9			0.110***	0.108***	0.146***
			(0.018)	(0.034)	(0.026)
Elder twin			0.034*	0.016	0.034*
			(0.020)	(0.042)	(0.019)
Heavier twin at birth			0.021	0.079	0.020
			(0.023)	(0.049)	(0.031)
Birthweight, gramms			0.000***	-0.000	0.000
			(0.000)	(0.000)	(0.000)
Constant	0.110***	0.007	-0.251***	0.082	-0.002
	(0.039)	(0.030)	(0.080)	(0.145)	(0.180)
Observations	3,877	3,877	3,877	1,186	3,877
R-squared	0.022	0.348	0.370	0.367	0.355
Twin FE	No	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	Yes	No

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table O10: The gender gap in teachers' assessments of children's mathematics abilities

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 3 OS subsample	(5) Model 4
Female	-0.224***	-0.120***	-0.123***	-0.147***	-0.159***
	(0.036)	(0.022)	(0.022)	(0.031)	(0.031)
Math level, age 9		0.804***	0.768***	0.761***	0.721***
		(0.012)	(0.013)	(0.024)	(0.020)
Verbal abilities, age 9			0.017	0.017	0.058***
			(0.013)	(0.023)	(0.020)
Non-verbal abilities, age 9			0.078***	0.115***	0.056***
			(0.013)	(0.023)	(0.018)
Elder twin			0.012	-0.015	0.019
			(0.015)	(0.030)	(0.016)
Heavier twin at birth			0.010	0.042	0.047*
			(0.018)	(0.035)	(0.026)
Birthweight, gramms			0.000	0.000	-0.000
			(0.000)	(0.000)	(0.000)
Constant	0.109***	-0.030	-0.084	-0.093	0.221
	(0.040)	(0.024)	(0.061)	(0.109)	(0.147)
Observations	3,877	3,877	3,877	1,186	3,877
R-squared	0.013	0.614	0.620	0.643	0.535
Twin FE	No	No	No	No	Yes
Cohort FE	Yes	Yes	Yes	Yes	No

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table O11: The role of parental and teachers' assessments in the gender gap in SAMA, FE models

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.218***	-0.005	-0.401***	-0.134	-0.210***	0.066
	(0.048)	(0.231)	(0.050)	(0.211)	(0.048)	(0.238)
Parental assessment of Math	0.591***	0.619***			0.564***	0.572***
	(0.039)	(0.050)			(0.040)	(0.054)
Female*parental assessment		-0.053				-0.015
		(0.054)				(0.065)
Teachers' assessment of Math			0.351***	0.392***	0.140***	0.173***
			(0.050)	(0.059)	(0.047)	(0.057)
Female*teachers' assessment				-0.079		-0.064
				(0.058)		(0.063)
Constant	-2.296***	-2.414***	-1.214***	-1.352***	-2.689***	-2.832***
	(0.254)	(0.296)	(0.278)	(0.306)	(0.274)	(0.316)
Observations	3,877	3,877	3,877	3,877	3,877	3,877
R-squared	0.297	0.298	0.188	0.189	0.301	0.301
Twin FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	No	No	No	No	No	No

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table O12: The role of the mathematics level of co-twin (CT) in the gender gap in SAMA

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.322***	-0.321***	-0.350***	-0.347***		
	(0.032)	(0.032)	(0.032)	(0.033)		
Math level of CT, age 9	-0.044**	-0.038	-0.049*	-0.080**	-0.087**	-0.024
	(0.018)	(0.024)	(0.025)	(0.038)	(0.040)	(0.033)
Female*Math level of CT		-0.013		0.047		
		(0.033)		(0.049)		
Has a male twin (MT)			-0.074**	-0.072**	-0.084*	-0.054
			(0.032)	(0.033)	(0.046)	(0.049)
MT*Math level of CT			0.014	0.066	0.064	-0.040
			(0.033)	(0.046)	(0.047)	(0.049)
Female*MT*Math level of CT				-0.103		
				(0.066)		
Constant	-0.065	-0.065	-0.017	-0.020	-0.083	-0.299**
	(0.089)	(0.089)	(0.091)	(0.091)	(0.124)	(0.123)
Observations	3,877	3,877	3,877	3,877	1,781	2,096
R-squared	0.175	0.175	0.176	0.177	0.187	0.116
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table O13: The role of co-twin (CT) SAMA - co-twin SAMA as binary variable

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.329***	-0.306***	-0.373***	-0.339***		
	(0.032)	(0.036)	(0.034)	(0.036)		
Has a male twin (MT)			-0.097***	-0.083**	-0.158***	0.002
			(0.037)	(0.036)	(0.051)	(0.055)
Confident twin	0.224***	0.287***	0.321***	0.168	0.127	0.403***
	(0.043)	(0.056)	(0.065)	(0.103)	(0.107)	(0.080)
Male twin*confident twin			-0.123	0.167	0.234*	-0.485***
			(0.081)	(0.118)	(0.128)	(0.118)
Female*confident twin		-0.122		0.209*		
		(0.078)		(0.127)		
Female*male twin*confident twin				-0.569***		
				(0.162)		
Constant	-0.101	-0.116	-0.042	-0.072	-0.070	-0.361***
	(0.089)	(0.090)	(0.091)	(0.091)	(0.122)	(0.123)
Observations	3,722	3,722	3,722	3,722	1,707	2,015
R-squared	0.181	0.182	0.185	0.188	0.198	0.129
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered within twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The variable "Confident twin" is a binary variable that equals 1 if co-twin's SAMA belongs to the top 20% of the SAMA distribution and 0 otherwise. Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight. CT refers to co-twins.

Table O14: The gender gap in self-assessed English abilities - the role of co-twins (CT)

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	0.248***	0.248***	0.297***	0.270***		
	(0.031)	(0.031)	(0.036)	(0.033)		
Has a male twin (MT)			0.127***	0.110***	0.109**	0.116***
			(0.036)	(0.033)	(0.050)	(0.041)
Self-assessed English ability of CT	0.181***	0.173***	0.223***	0.074	0.071	0.273***
	(0.024)	(0.034)	(0.033)	(0.049)	(0.049)	(0.040)
Female*self-assessed English ability of CT		0.015		0.193***		
		(0.042)		(0.063)		
MT*self-assessed English ability of CT			-0.070*	0.142**	0.139**	-0.236***
			(0.042)	(0.065)	(0.065)	(0.051)
Female*MT*self-assessed English ability of CT				-0.376***		
				(0.098)		
Constant	-0.140*	-0.139*	-0.224***	-0.196**	-0.201	0.084
	(0.078)	(0.078)	(0.082)	(0.080)	(0.124)	(0.094)
Observations	3,876	3,876	3,876	3,876	1,781	2,095
R-squared	0.112	0.112	0.116	0.123	0.114	0.113
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p«0.01, ** p«0.05, * p«0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table O15: The gender gap in self-assessed physical abilities - the role of co-twins (CT)

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 4 boys	(6) Model 4 girls
Female	-0.058*	-0.058*	-0.055	-0.048		
	(0.031)	(0.031)	(0.037)	(0.033)		
Has a male twin (MT)			0.009	0.009	0.014	-0.004
			(0.038)	(0.033)	(0.047)	(0.045)
Self-assessed physical ability of CT	0.236***	0.230***	0.246***	0.083*	0.085*	0.305***
	(0.027)	(0.037)	(0.033)	(0.050)	(0.050)	(0.040)
Female*self-assessed physical ability of CT		0.012		0.223***		
		(0.044)		(0.064)		
MT*self-assessed physical ability of CT			-0.021	0.220***	0.218***	-0.230***
			(0.045)	(0.068)	(0.068)	(0.064)
Female*MT*self-assessed physical ability of CT				-0.453***		
				(0.116)		
Constant	-0.107	-0.107	-0.111	-0.117	-0.070	-0.205**
	(0.076)	(0.076)	(0.082)	(0.080)	(0.123)	(0.093)
Observations	3,853	3,853	3,853	3,853	1,767	2,086
R-squared	0.062	0.062	0.062	0.073	0.065	0.079
Twin FE	No	No	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes

Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table O16: The role of parental education, age nine

VARIABLES	(1) SAMA	(2) SAMA	(3) SAMA	(4) Math level	(5) Math level	(6) Math level
Female	-0.327***	-0.183*	-0.274*	-0.128***	0.113	0.047
	(0.032)	(0.101)	(0.147)	(0.034)	(0.109)	(0.142)
Parental education						
High-grade CSE/GCSE	-0.128**	-0.053		0.189***	0.339***	
	(0.060)	(0.086)		(0.066)	(0.101)	
A-level or below degree	-0.090	-0.002		0.373***	0.542***	
	(0.061)	(0.087)		(0.067)	(0.102)	
Degree	-0.129**	-0.032		0.667***	0.809***	
	(0.061)	(0.088)		(0.066)	(0.099)	
Female*High-grade CSE/GCSE		-0.136	-0.191		-0.265**	-0.222
		(0.116)	(0.179)		(0.125)	(0.167)
Female*A-level or below degree		-0.167	-0.217		-0.298**	-0.274*
		(0.116)	(0.178)		(0.127)	(0.162)
Female*Degree		-0.178	-0.188		-0.247**	-0.254
		(0.115)	(0.168)		(0.124)	(0.161)
Math level, age 9						
Constant	0.070	-0.045	0.042	-0.179***	-0.315***	0.184***
	(0.098)	(0.114)	(0.237)	(0.066)	(0.090)	(0.024)
Observations	3,863	3,863	3,863	3,863	3,863	3,863
R-squared	0.173	0.175	0.163	0.060	0.062	0.013
Twin FE	No	No	Yes	No	No	Yes
Cohort FE	Yes	Yes	No	Yes	Yes	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table O17: The role of maternal characteristics in the gender gap in SAMA, age nine

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
Female	-0.290***	-0.320***	-0.318***	-0.385***	-0.437***	-0.422***
	(0.042)	(0.033)	(0.037)	(0.070)	(0.055)	(0.061)
Mother has A-levels or above	0.016					
	(0.046)					
Female*Mother has A-levels or above	-0.081			-0.145		
	(0.063)			(0.101)		
Mother has managerial job		0.002				
		(0.072)				
Female*Mother has managerial job		-0.039			-0.083	
		(0.105)			(0.150)	
Mother needs qualification			-0.027			
			(0.050)			
Female*Mother needs qualification			-0.030			-0.101
			(0.071)			(0.109)
Constant	-0.074	-0.066	-0.058	0.027	0.034	0.033
	(0.090)	(0.089)	(0.089)	(0.236)	(0.236)	(0.236)
Observations	3,877	3,877	3,877	3,877	3,877	3,877
R-squared	0.174	0.174	0.174	0.165	0.164	0.164
Twin FE	No	No	No	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	No	No	No

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.

Table O18: The role of stereotypical parental assessments in the twin peer effects in SAMA, age nine

VARIABLES	(1) Not OE boys	(2) OE boys	(3) Not UE girls	(4) UE girls
Has a male twin (MT)	-0.074	-0.017	-0.043	-0.251***
	(0.054)	(0.067)	(0.055)	(0.083)
SAMA of CT, age 9, std	0.046	0.040	0.209***	0.245***
	(0.048)	(0.051)	(0.041)	(0.063)
MT*SAMA of CT	0.167**	-0.079	-0.198***	-0.180*
	(0.065)	(0.070)	(0.064)	(0.097)
Constant	-0.170	0.535***	-0.169	-0.381*
	(0.131)	(0.195)	(0.126)	(0.218)
Observations	1,256	451	1,489	526
R-squared	0.270	0.099	0.214	0.126
Twin FE	No	No	No	No
Cohort FE	Yes	Yes	Yes	Yes

Notes: Source: TEDS (Rimfeld et al., 2019). Robust standard errors clustered by twin pairs in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Further control variables: mathematics level at age nine, verbal and non-verbal cognitive skills at age nine, elder twin, heavier twin, and birth weight.