Biology and the gender gap in educational performance - The role of prenatal testosterone in test scores[*]

Anne C. Gielen[†] Esmée S. Zwiers[‡]

January 24, 2018

WORK IN PROGRESS, PLEASE DO NOT CITE OR REDISTRIBUTE

Abstract

Explanations for gender differences in educational performance are generally sought in gender-biased environments and socialization. We investigate whether biology, and specifically prenatal testosterone, can explain these differences. Prenatal testosterone is elevated for individuals with a male co-twin due to testosterone transfers in utero. Given that the sex of the co-twin is determined randomly, this provides exogenous variation in prenatal testosterone. Administrative data on a large sample of Dutch twins is used. After controlling for the effects of growing up with a same-sex or opposite-sex sibling, we find null-effects for boys with a male co-twin on three measures of educational performance (aggregate ability, reading and math). Girls with a male co-twin score lower on math by 7% of a standard deviation. This is mainly caused by more girls ending up in the bottom 10% of the math-score distribution. No effects are found on a reading and aggregate ability score.

Keywords: gender performance gap, twins, testosterone
JEL: I20, J16

[*] The authors wish to thank Thomas Buser, Henrik Cronqvist, Gordon Dahl, Jan Kabátek, Guido Imbens, Olivier Marie, David Neumark, Hannes Schwandt, Zahra Siddique, Andrea Terei, Dinand Webbink, Bas ter Weel, Basit Zafar, and seminar participants in the TI BrownBag Seminar Series, the TI PhD Lunch Seminar Series, the annual TI PhD Jamboree, the 31^{st} annual ESPE Conference, the 32^{nd} annual EEA Conference, and the Netherlands Economists Day 2017 for valuable comments.

[†] Department of Economics, Erasmus University Rotterdam, and Tinbergen Institute, and IZA. E-mail: gielen@ese.eur.nl

[‡] Department of Economics, Erasmus University Rotterdam, and Tinbergen Institute. E-mail: zwiers@ese.eur.nl

1 Introduction

Gender differences in math and language test scores have been long known and appear to be rather persistent. Generally, boys outperform girls in mathematics (Fryer and Levitt, 2010; Bharadwaj et al., 2015), particularly among children at the high end of the ability distribution (Ellison and Swanson, 2009; Stoet and Geary, 2013; Pope and Sydnor, 2010), but in the reading domain girls outperform boys (Halpern et al., 2007; Guiso et al., 2008; Banda et al., 2010). These differences are important as they may explain gender differences in educational and occupational choices in adulthood (Buser et al., 2014; Banda et al., 2010; Ceci et al., 2009), as well as gender-related earnings differentials[1]. The existing literature has shown that gender differences in math and reading ability arise from social conditioning and gender-biased environments (Wilder and Powell, 1989; Miller and Halpern, 2014), yet little is known to what extent biological factors are an important driver of these gender differences in test scores. A role for biological factors in creating these gender-specific outcomes may imply that we are currently over-estimating the role of any discriminatory or gender-biased factors.

This paper explores the role of biology in explaining the gender gap in math and reading test scores in childhood, and focuses in particular on the role of prenatal testosterone. Prenatal testosterone induces the sexual differentiation of the male fetus. In addition to influencing the development of sexually dimorphic physical characteristics, exposure to prenatal testosterone is known to wire the brain with masculine behavioral patterns (i.e. in preferences, personality, and temperament) (Jordan-Young, 2010). Evidence from laboratory and field experiments indicates that women display less aggressive behavior (e.g. Bettencourt and Miller (1996)), act more risk averse (e.g. Eckel and Grossman (2008); Croson and Gneezy (2009)), and engage less in competitive activities (e.g. Gneezy et al. (2003); Niederle and Vesterlund (2007); Buser (2012b); Örs et al. (2013)) than men. Little is known to what extent these differences are explained by biology or socialization, and to what extent they translate into gender-specific primary school outcomes such as math and reading test scores.

In this paper, we exploit a natural experiment in twinning to identify the biological contribution of prenatal testosterone exposure to gender differences in test scores. Since it is impossible to directly measure and link prenatal testosterone exposure to test score performance in primary school, we exploit twin testosterone transfers (TTT) as an exogenous proxy. Between the eighth and twenty-fourth week of gestation male fetuses are exposed to elevated levels of testosterone (Auyeung et al., 2013). As with other litter-bearing mammals, among human twins this testosterone can

[1] For an overview of the literature, trends and explanations of the gender pay gap consult Blau and Kahn (2000), and Blau and Kahn (2016).

transfer in significant concentrations from a male twin to his female uterus mate. Previous studies using twin testosterone transfers (TTT) suggest that females with a fraternal co-twin are more masculine in morphological characteristics, behavior, and cognitive capabilities (Resnick et al., 1993; Cohen-Bendahan et al., 2004; Peper et al., 2009; Vuoksimaa et al., 2010a,b; Heil et al., 2011; Slutske et al., 2011). For males with a male co-twin no increased masculine behavior or characteristics are found (Resnick et al., 1993; Peper et al., 2009; Tapp et al., 2011; Cronqvist et al., 2015). In this paper, we argue that twinning is a plausible natural experiment to measure the effect of elevated prenatal testosterone concentrations on math and reading test scores.

Earlier applications of twin testosterone transfers (TTT) to economic outcomes are relatively scarce. A study by Gielen et al. (2016) investigates the role of TTT to explain the gender wage gap, and finds higher earnings for men with a male co-twin, but no effect for women. Another study by Cronqvist et al. (2015) focuses on financial decision-making, and finds that females with a male co-twin take significantly more risk later in life compared to females with a female co-twin. This paper is the first to study the role of TTT in gender differences in educational performance. We use Dutch administrative data from Statistics Netherlands where we observe test-score data of nearly 76,416 twins born between 1993 and 2003, of which 39,441 individuals can be matched to test-score records. These data allow us to estimate the effect of having a male co-twin on math and reading test scores in the final grade of primary education (i.e. at approximately age twelve) in the years 2006 to 2014.

Our results suggest that prenatal testosterone does not affect performance for boys (on an aggregate, reading and math score). However for girls we find that prenatal testosterone decreases math scores by 7% of a standard deviation, whereas null effects are found on an aggregate and a reading score. This effect is likely driven by the fact that girls exposed to higher levels of prenatal testosterone are 2.5% more likely to end up in the bottom 10% of the math-score distribution. This result is counterintuitive as one would expect that girls with more prenatal testosterone are more male-typical and hence would perform better at math. A potential explanation is that girls with a male co-twin are more male-typical in morphological characteristics and behavior which might interact negatively with educational outcomes at age twelve.

The structure of this paper is as follows. The next section provides background information on the gender gap in math and reading test scores, and the potential role of prenatal testosterone. Section 3 outlines the identification strategy, and the data and results are presented in sections 4 and 5. These are followed by a discussion (section 6) and conclusion (section 7).

2 Prenatal testosterone and the gender math gap

Boys on average perform better at math than girls in a majority of countries (Fryer and Levitt, 2010; Banda et al., 2010; Bharadwaj et al., 2015; OECD, 2015). The gap widens with age (Fryer and Levitt, 2010; Bharadwaj et al., 2015), and ability (Ellison and Swanson, 2009; Fryer and Levitt, 2010; Pope and Sydnor, 2010; Stoet and Geary, 2013; OECD, 2015). Ellison and Swanson (2009). The math differential is reversed in the reading domain, where girls generally outperform boys (Halpern et al., 2007; Guiso et al., 2008; Banda et al., 2010). Apart from higher average performance on math, and lower average performance on reading, boys are also known to be more variable in performance (Halpern et al., 2007; Machin and Pekkarinen, 2008). The latter implies that boys are more often in both the high and low ends of the performance distribution.

Gender differences in educational performance are attributed to either (1) biological differences as discrepancies in brain development or testosterone, or (2) gender differences in socialization, stereotypes, and preferences (Wilder and Powell, 1989; Miller and Halpern, 2014). The existing literature examines explanations for the latter channel, e.g.: differences in the cultural dimension (Guiso et al., 2008; Stoet and Geary, 2013), gender differences in competitiveness (Gneezy et al., 2003; Gneezy and Rustichini, 2004; Niederle and Vesterlund, 2007; Croson and Gneezy, 2009; Flory et al., 2010; Niederle and Vesterlund, 2010; Buser, 2012b; Örs et al., 2013), stereotype threat (e.g. Spencer et al. (1999); Stoet and Geary (2012); Nollenberger et al. (2014)), and gender biased environments (Fryer and Levitt, 2010; Bharadwaj et al., 2015). Little is known about biological determinants of gender differences in educational performance.

It is well known that the early life environment is important for the development of a child's cognitive capacities (e.g. Carneiro and Heckman (2003); Knudsen et al. (2006); Heckman (2008); Currie and Almond (2011)). The pre-birth environment plays a big role alongside the post-birth environment. The fetal origins hypothesis asserts that the prenatal period is of crucial importance for both the cognitive development and the health of the child. The fetus is very sensitive to -amongst others- smoking during pregnancy, maternal malnutrition, and maternal stress, and these factors can have large impacts long after birth (e.g. Almond and Currie (2011); Scholte et al. (2015)). This paper will consider the impact of a prenatal hormonal factor in influencing cognitive capacities: prenatal testosterone.

2.1 The role of prenatal testosterone

We examine the role of prenatal factors, and specifically prenatal testosterone, in determining gender differences in educational performance. Testosterone is the main androgen causing sexual differentiation of the male fetus. Males experience three periods of elevated testosterone exposure, whereas female testosterone levels remain rather constant over the life-cycle. These critical periods are between the eighth and twenty-fourth week of gestation (prenatal testosterone surge which causes sexual differentiation of the fetus), three to four months after birth, and in puberty Auyeung et al. (2013).

Prenatal testosterone production starts at around the seventh and eighth week of gestation and continuous until approximately week twenty-four. Prenatal testosterone is responsible for the development of the testes (Tapp et al., 2011), but this period of gonadal development would also be critical for the development of the fetal brain (Van de Beek et al., 2004).[2] Prenatal testosterone would wire the brain with masculine behavioral patterns (i.e. in preferences, personality, and temperament) (Jordan-Young, 2010). The female fetus is exposed to much lower levels of prenatal testosterone (Tapp et al., 2011; Auyeung et al., 2013).

2.1.1 Proxies for prenatal testosterone

The best measure for prenatal testosterone is fetal serum, which is unfeasible due to the risks associated with its collection Other proxies, like maternal serum testosterone, umbilical cord serum, and amniotic fluid concentrations all have their own disadvantages (Van de Beek et al., 2004). Earlier studies used medical conditions and 2D:4D digit ratios as proxies for prenatal testosterone. Clinical studies examine the effects of prenatal testosterone exposure on cognitive ability by studying women subject to congenital adrenal hyperplasia (CAH). Females with this condition are prenatally exposed to high levels of androgens (Speiser and White, 2003). To illustrate, women diagnosed with CAH are found to perform better on spatial tasks than control women (Puts et al., 2008). Disadvantages of using clinical samples are the usually small sample size, and limited external validity(Baron-Cohen et al., 2004).

The 2D:4D ratio (the ratio of the index to the ring finger) is regarded as a (noisy) marker for prenatal testosterone (Cohen-Bendahan et al., 2005). The ratio is sexually dimorphic as it is, on average, lower for men than for women (Lutchmaya et al., 2004; Medland et al., 2008). Elevated fetal testosterone levels are associated with lower 2D:4D ratios (Lutchmaya et al., 2004), and girls diagnosed with CAH are found to have lower 2D:4D ratios (Puts et al., 2008).

Lower 2D:4D ratios would be associated with lower risk-averseness (Dreber and Hoffman, 2007; Coates et al., 2009; Garbarino et al., 2011), aggressiveness and increased sensation-seeking (Hampson et al., 2008),

[2] Sexual differentiation of the brain is said to take place between the 14th and 19th week of gestation (Baron-Cohen et al., 2004).

more male-typical preferences in occupational choices for women (Nye and Orel, 2015), social preferences (Buser, 2012a), better performance in sports (Manning and Taylor, 2001), and an elevated physical fitness (Hönekopp et al., 2007). Lower 2D:4D ratios are positively correlated with performance on mental rotation tasks (Manning and Taylor, 2001), whereas this relationship is not confirmed by Austin et al. (2002) and Coolican and Peters (2003). The 2D:4D ratio is considered as a proxy for prenatal testosterone, although it is considered a noisy biomarker as digit ratios would be more correlated with ethnicity than with sex (Cohen-Bendahan et al., 2005).

2.1.2 Twin testosterone transfers

Recently prenatal testosterone exposure is studied using twins. Individuals with a male co-twin would be exposed to high levels of prenatal androgens. This can be exploited as a natural experiment -given that the sex of the co-twin is random (Tapp et al., 2011). The transferring of testosterone across amniotic membranes during gestation is better known as twin testosterone transfers (TTT). Babies with a male co-twin would be exposed to higher levels of prenatal androgens during gestation through in-utero testosterone transfers.

The existence of twin testosterone transfers was first documented in animal-studies, where female rodents with a position near their brothers in the womb would display more male-typical behavior (for an overview see Cohen-Bendahan et al. (2005)). The existence of a similar channel for humans is documented by Miller (1994). Direct testing of twin testosterone transfers is very difficult. Animal studies (i.e. with rodents) allow for a direct manipulation of prenatal testosterone levels, which is unethical for humans (Cohen-Bendahan et al., 2005). Twin studies show that females with a male co-twin have a more masculine brain structure (Cohen-Bendahan et al., 2004) and volume (Peper et al., 2009), are more likely to be right-handed which is an indicator of high exposure to testosterone (Vuoksimaa et al., 2010a), do better at mental rotation tasks than females with a female co-twin (Vuoksimaa et al., 2010b; Heil et al., 2011), and are more sensation-seeking (Resnick et al., 1993; Slutske et al., 2011). Studies investigating digit ratios in relationship to twin testosterone transfers found lower 2D:4D ratios for opposite sex twin females (van Anders et al., 2006; Voracek and Dressler, 2007), although this result is not confirmed by Medland et al. (2008).

Some studies fail to find effects for males with a male co-twin even though these males should also be exposed to higher levels of prenatal testosterone (Resnick et al., 1993; Peper et al., 2009; Tapp et al., 2011; Cronqvist et al., 2015). Tapp et al. (2011) argue that the effect is likely less obvious for males, as males themselves are already exposed to relatively high levels of prenatal testosterone.

We use twin testosterone transfers as a proxy for fetal testosterone. There are two earlier applications of TTT within economics. Gielen et al. (2016) use TTT to examine the influence of testosterone on the gender wage gap. Although positive effects of prenatal testosterone exposure are found for men, prenatal testosterone is not associated with increased earnings for women. Cronqvist et al. (2015) use TTT to explain gender differences in financial decision making and find that higher exposure to prenatal testosterone can explain masculinization of investing behavior, implying that females with a fraternal male co-twin undertake more risky investments. This is the first application of TTT to gender differences in educational performance.

3 Empirical strategy

This paper exploits twinning to examine the effects of prenatal testosterone on test scores. An individual with a male co-twin is exposed to higher concentrations of fetal testosterone due to testosterone transfers. Three assumptions must hold for establishing the causal effect of prenatal testosterone on test scores, namely: (1) there are testosterone transfers from a male fetus to the neighboring fetus, (2) the distribution of sexes is random among twin pairs, and (3) there are no confounding factors related to opposite sex twinning which can affect educational outcomes of children via other routes than testosterone transfers.

Direct tests of the first assumption in humans are not available to our knowledge. However, direct testing on animals showed that in-utero testosterone transfer exist (see Cohen-Bendahan et al. (2005) for an overview). The channel was extended to humans by Miller (1994), and ever since has been supported by indirect evidence linking twinning to testosterone transfers. Multiple studies report increased masculine morphological, cognitive and behavioral characteristics for women with a fraternal male co-twin (Resnick et al., 1993; Cohen-Bendahan et al., 2004; Peper et al., 2009; Vuoksimaa et al., 2010a,b; Heil et al., 2011; Slutske et al., 2011). No effects of prenatal testosterone are found for males with a male co-twin, possibly due to their already high exposure to prenatal testosterone (Resnick et al., 1993; Peper et al., 2009; Tapp et al., 2011; Cronqvist et al., 2015). Tapp et al. (2011) conclude that the evidence on TTT is incomplete, but it is sufficient to authorize further investigations.

The second identifying assumption is that the distribution of sexes is random among twin pairs. Implying that whether an individual has a twin-brother or sister is randomly determined and not influenced by confounders that can also influence the outcome variables. Twins can be monozygotic, when one fertilized egg splits into two fetuses, or dizygotic, when two fertilized eggs develop into fetuses. Monozygotic (identical) twin pairs are always same-sex, whereas dizygotic (fraternal) twins can be same-sex or opposite-sex.

The sex of a child is depending on whether the male's fertilizing spermatozoon carries an X or a Y chromosome, which is regarded as random. Despite it, human sex ratios[3] are weighted towards boys as it is 105.9 for singletons and 103.2 for twins. Several theories for the discrepancy in sex rates between singletons and twins include differences in gonadotrophin levels (hormone responsible for reproductive functioning) at time of conception, higher mortality for male fetuses in twin pairs, and differences by race (Fellman and Eriksson, 2010)

Identical twins generally have lower sex ratios than fraternal twins[4], which is due to an anomaly which is inherent in X-chromosomes which makes them more likely to divide, and hence form a identical twin pair. On the contrary, fraternal twins are more likely to be male, which is likely due to higher maternal levels of steroid hormones (testosterone and estrogen) at conception (James, 2010). Maternal serum testosterone levels are not a good proxy for actual prenatal testosterone (Van de Beek et al., 2004; Cohen-Bendahan et al., 2005), which might suggest that maternal testosterone does not affect prenatal testosterone. If maternal and fetal testosterone levels interact it would only strengthen the identification as individuals with a male co-twin would be exposed to even higher levels of prenatal testosterone (Gielen et al., 2016).

Thirdly, the sex of the co-twin cannot be related to educational outcomes through ways other than testosterone transfers. This assumption is likely violated as growing up with a brother is different from growing up with a sister, and the sibling's gender may eventually affect educational outcomes. Comparing the educational outcomes of a twin with a male uterus mate and a twin with a female uterus mate will measure the effect of prenatal testosterone, but additionally the effect of growing up with a same-sex or opposite-sex sibling. To isolate the effect of prenatal testosterone a control group of closely spaced singletons (CSS) is used.[5] The effect of prenatal testosterone is isolated if sibling socialization is similar for twins and this group of CSS.

This control group allows us to disentangle the effect of prenatal testosterone from the combined effect of prenatal testosterone and socialization, but it also imposes two extra assumptions on the identification strategy. First, socialization must be similar for twins and closely spaced singletons (CSS). The close spacing of the control group makes sure that it is likely that siblings born very near each other experience similar environments.

[3] Sex ratios represent the number of boys born for every one hundred girls.

[4] Gielen et al. (2016) find a sex ratio of 94.2 using data from James (2010).

[5] This approach is suggested by Cohen-Bendahan et al. (2005) and Tapp et al. (2011) and employed by Gielen et al. (2016).

Hence interactions are likely more twin-like than for regular siblings (who are born more than twelve months apart). We execute a robustness check to assert that this concern does not affect our results.

Second, the child's level of prenatal testosterone must be independent of whether the child has a closely spaced brother or sister. We know that prenatal testosterone in male singletons declines with birth order (as measured by umbilical cord serum) when spacing between children is less than four years (Maccoby et al., 1979; Baron-Cohen et al., 2004). This implies that it might be that within a closely spaced singleton pair, especially the male child might have experienced lower levels of prenatal testosterone. We estimate the model using only first-borns to assert that this potential concern does not play a role.

Another assumption we need to make is that socialization is similar for identical and fraternal twins as we cannot distinguish zygosity. This brings us to the equal environments assumption, which states that there are no systematic differences in the upbringing environments of identical and fraternal twins (Gielen et al., 2016). We assume that there are no differences in the upbringing environments of identical and fraternal twins that can affect educational outcomes at age 12. Obviously, there might be differences between identical and fraternal twins, especially as identical twins share 100% of their genetic material, whereas this is approximately 50% for fraternal twins. However, the EEA is not violated in several areas (Matheny et al., 1976; Scarr and Carter-Saltzman, 1979; Kendler et al., 1994; Hettema et al., 1995; Eriksson et al., 2006; LoParo and Waldman, 2014), and most importantly for spatial ability Derks et al. (2006).

A last assumption we have to make is that factors that determine whether a child is in a twin-pair or in a closely spaced singleton pair are unrelated to educational outcomes. This assumption is likely violated as twins and closely spaced singletons are born into different families, which are likely not random draws from the population. A rich set of control variables is used to take into account these differences. We apply propensity score matching to make the sample of twins and CSS more comparable. A more detailed explanation of the respective differences and used controls can be found in section 4.3.

Specification (1) is estimated for a sample of twins and closely spaced singletons, for different outcome variables (y_i), and is used to identify the effect of TTT. The outcome variables are an aggregate test-score, and sub-scores in the domains of math and reading. Equation (1) contains a female indicator, an indicator for being part of a twin-pair, an indicator for being part of an opposite-sex sibling pair, and their respective interactions. Vector $\mathbf{X}_i$ contains control variables, and u_i is the individual-specific error term. Standard errors are clustered on the maternal identification number.

$$
\begin{aligned}
y_i = \beta_0 + \beta_1 female_i + \beta_2 OS_i + \\
\beta_3 twin_i + \beta_4 (female_i OS_i) + \beta_5 (twin_i female_i) + \\
\beta_6 (twin_i OS_i) + \beta_7 (female_i OS_i Twin_i) + \mathbf{X}_i \delta + u_i
\end{aligned} \tag{1}
$$

For a boy the effect of having a twin brother is captured by $-\beta_2 - \beta_6$, and the socialization effect of having a brother is captured by $-\beta_2$. Implying that the effect of TTT for males is shown by $-\beta_6$. For girls the effect of having a male twin are captured by $\beta_2 + \beta_4 + \beta_6 + \beta_7$, and the socialization effect of having a brother is entailed in $\beta_2 + \beta_4$. The effect of TTT for girls is $\beta_6 + \beta_7$.

4 Data

4.1 Dutch twins

Dutch administrative date is obtained from Statistics Netherlands.[6] Individuals can be matched across datasets with a Random Identification Number (RIN). The Parent-Child data is used, which matches children to any parent alive between 1995 and 2015, to compile a dataset of Dutch twins. It contains information on $15,860,240$ individuals. We drop stillbirths ($n = 22,290$) and individuals with missing RIN ($n = 547,350$).

Table 1: Frequency of family structures in 2015 GBA

Family type	Frequency	Percent
Only child	214,509	9.16
Singleton (closest sibling > 12 months)	2,020,799	86.29
Singleton (closest sibling ≤ 12 months)	27,628	1.18
Twin	76,416	3.26
Higher order multiple	2,462	0.11
Total	2,341,814	100.00

Notes: Frequency of family structures for individuals born 1993-2003 (this is the time-period in which children are born for whom we observe educational outcomes), whose mother can be identified in the data, and who have less than 15 siblings through either parent.

This data is supplemented with demographic characteristics from the Municipal Population dataset (Gemeentelijke Basisadministratie). It contains information on the individuals' year and month of birth, the parents' year and month of birth, sex, country of origin. We identify closely spaced

[6] Statistics Netherlands provides non-public microdata which can be accessed remote-access after signing a confidentiality agreement.

singletons as siblings whose birth dates are at most 12 months apart. Individuals with more than 15 siblings through either parent are dropped from the sample ($n = 2,090$).

The distribution of family structures for the remaining sample is shown in Table 1. The twinning probability (3.26%) is consistent with the incidence of twinning in the Netherlands between 1993 and 2004 (3.39%).[7] Children without siblings, with siblings born outside the 12 month range, and higher order multiples are dropped from the sample, which leaves a sample of twins and CSS.

Sibling pairs in the remaining sample are identified as same-sex or opposite-sex siblings. We drop individual cases if it is difficult to determine the sex composition, e.g. when there are three CSS in one family (small fraction of 4.5%). Closely spaced singletons whose birth dates are within 7 months are dropped from the sample ($n = 17,462$). The distribution of twins and CSS by gender composition is shown in Table 2.[8]

Table 2: Twins and closely spaced singletons

	Observed in GBA		Observed in Test Score Data	
	Frequency	Percent	Frequency	Percent
Females				
OS Twin	13,626	13.4	7,608	14.9
SS Twin	24,222	23.7	12,601	24.7
OS CSS	6,457	6.3	2,995	5.9
SS CSS	6,015	5.9	2,839	5.6
Males				
OS Twin	13,626	13.4	7,193	14.1
SS Twin	24,942	24.4	12,039	23.6
OS CSS	6,415	6.3	2,805	5.5
SS CSS	6,730	6.6	2,886	5.7
Total	102,033	100.00	50,966	100.00

Notes: The first column shows the distribution of opposite-sex and same-sex pairs in the GBA (1993-2003). The second panel shows the same distributions for the test score data.

[7] Authors' calculations based on birth figures available (online) at Statistics Netherlands. This number is upward biased as it does not take into account stillbirths.

[8] The twins-sample contains 65.7% same-sex and 34.3% opposite-sex pairs born from 1993 to 2003. The number of dizygotic twins can be approximated as twice the number of opposite-sex twins according to Weinberg's differential method (for empirical tests see Vlietinck et al. (1988) and Fellman and Eriksson (2006)), implying that approximately 68.6% of the twins in our sample are dizygotic.

4.2 Educational outcomes

Data on primary school test-scores is also provided by Statistics Netherlands. The data contains information on a standardized test performed in the eight and final grade of elementary education (Cito-test). We use data for the years 2006 to 2014[9], which is available for individuals attending schools who gave permission to transfer test-scores to Statistics Netherlands. Children without identification number, and whose monther's age at birth is missing ($n = 62,293$) are dropped from the sample. The latest score is preserved for children having multiple test-score records in the data. When we merge the test-score data with the demographics a sample containing $50,966$ individuals remains.

The standardized test incorporates language, math, information processing, and world orientation. The latter is optional and hence not completed by all children. The scores on the remaining parts are translated in an aggregated score ranging between 501 and 550. The aggregate score is standardized for interpretation purposes. We use Z-scores for math and reading (which are standardized by year).

4.3 Descriptive statistics

Twins and closely spaced singletons are likely born into different families, which is shown in Table 3. This table also clearly shows that twins and CSS are different from the full population.

Twins are born to older mothers, as twinning probabilities increase with maternal age (Rosenzweig and Wolpin, 1980a; Bronars and Grogger, 1994; Jacobsen et al., 1999), the use of artificial reproductive technologies (ART) (Bhalotra et al., 2016), and parity (Rosenzweig and Wolpin, 1980a,b). This also explains that twins have a lower birth order on average. Twins are more often born into 2-parent households, and these households have a higher earnings capacity as measured by household income and the labor market status of the mother.[10]

The age at taking the test is higher, on average, for twins and CSS as compared to the full population. Twin pregnancies are considered risky, and it is not uncommon for twins to be born after shorter gestation than singletons (Almond et al., 2005; Bhalotra et al., 2016). Shorter gestational duration can disadvantage twins throughout their life which might result in a higher age at the time of taking the test. CSS are likely born in low

[9] Test scores for 2015 are available but are not used as the set-up of the test changed in 2015.

[10] Household income is measured in the year the child turns four. Income information is available from 1999, implying that we have this information for all children born after 1994. Household income is compiled using income from employment and income from self-employment. Household income is the sum of the earnings of both parents in a particular year (age at birth or age at birth plus four) and is corrected for inflation.

socio-economic status families which can explain their higher average age at taking the test.

Table 3: Descriptive statistics

		Female twins and closely spaced singletons							
		OS Twin	SS Twin	OS CSS	SS CSS	All females	Twin -		
		(1)	(2)	(3)	(4)	(5)	CSS	1-2	3-4
Variable	*Def.*	*n=7,608*	*n=12,601*	*n=2,995*	*n=2,839*	*n=641,882*			
Total score	Std	-0.0883	-0.0553	-0.238	-0.253	-0.009	***	**	
Language	Std	0.0569	0.0655	-0.128	-0.147	0.124	***		
Math	Std	-0.236	-0.178	-0.299	-0.304	-0.157	***	***	
Age	Months	12.048	12.048	12.073	12.092	11.982	***		
Birth order		1.735	1.743	2.106	2.130	1.806	***		
Spacing		0	0	11.483	11.490		***		
Nonnative	I	0.158	0.166	0.382	0.421	0.211	***		***
Family size	Via mother	2.986	3.058	3.475	3.593	2.601	***	***	***
Mother's age	At birth	31.991	31.356	28.949	28.374	30.529	***	***	***
Father's age	At birth	34.632	33.935	32.406	32.091	33.313	***	***	**
Mother in DI	I(in DI)	0.0201	0.0162	0.0190	0.0155	0.0129		**	
HH-type	2-parent	85.66	85.52	80.63	79.36	84.81	***		
	1-parent	13.93	13.88	17.93	19.20	14.75			
	Other	0.29	0.49	1.20	1.34	0.33			
	Missing	0.12	0.11	0.23	0.11	0.11			
		n=6,552	*n=10,660*	*n=2,513*	*n=2,314*	*n=543,672*			
HH-income*	Child=4	44,023.21	43,014.93	32,906.84	31,706.77	41,144.33	***	*	
Mother* working	I(works)	0.634	0.635	0.476	0.471	0.635	***		

		Male twins and closely spaced singletons							
		OS Twin	SS Twin	OS CSS	SS CSS	All males	Twin-		
		(1)	(2)	(3)	(4)	(5)	CSS	1-2	3-4
Variable	*Def.*	*n=7,193*	*n=12,039*	*n=2,805*	*n=2,886*	*n=636,303*			
Total score	Std	0.0419	0.0373	-0.188	-0.189	0.0393	***		
Language	Std	-0.0708	-0.0775	-0.356	-0.351	-0.0792	***		
Math	Std	0.174	0.185	0.0706	0.0735	0.185	***		
Age	In months	12.067	12.108	12.125	12.114	12.037	***	***	
Birth order		1.730	1.756	2.138	2.137	1.805	***	*	
Spacing		0	0	11.481	11.490		***		
Nonnative	I	0.158	0.173	0.397	0.372	0.210	***	***	*
Family size	Via mother	2.974	3.068	3.491	3.519	2.597	***	***	
Mother's age	At birth	32.008	31.497	28.920	28.702	30.568	***	***	
Father's age	At birth	34.637	34.065	32.395	32.400	33.309	***	***	
Mother in DI	I(in DI)	0.0196	0.0161	0.0175	0.0144	0.0121		*	*
HH-type	2-parent	85.97	85.98	80.46	79.49	85.18	***		
	1-parent	13.69	13.53	17.83	19.44	14.41			
	Other	0.22	0.37	1.50	1.04	0.30			
	Missing	0.11	0.12	0.21	0.03	0.11			
		n=6,147	*n=10,151*	*n=2,315*	*n=2,417*	*n=535,643*			
HH-income*	Child=4	44,973.46	43,344.22	32,484.99	33,062.50	41,610.28	***		
Mother* working	I(works)	0.642	0.646	0.475	0.492	0.641	***		

[*] Lower number of observations as data is available for children born after 1994.

[**] Notes: The reported means are presented for the sample which is discussed in more detail in section three.

Holding gender constant, there are not many significant differences in test scores between opposite-sex and same-sex siblings. However, females in opposite-sex twin pairs score significantly lower at math and the aggregate score as opposed to same-sex twin girls. This is suggestive evidence for TTT affecting girls' math and aggregate score negatively. This simple comparison neglects potential socialization effects causing educational differentiation. We need the control sample of closely spaced singletons to say more about the effects of prenatal testosterone.

There are other significant differences between opposite-sex and same-sex siblings pairs. Opposite-sex sibling pairs generally have smaller families, and older parents at giving birth. This could hint at a parental preference for children of mixed genders (e.g. Angrist and Evans (1998)).

Table 4: Gender gaps in test performance

		All			Sample	
	Boys	Girls	Δ	Boys	Girls	Δ
Score	n=636, 303	n=641, 882		n=24, 923	n=26, 043	
Total	0.039	-0.009	0.05***	-0.013	-0.107	0.09***
Reading	-0.079	0.124	-0.20***	-0.138	0.018	-0.16***
Math	0.185	-0.157	0.34***	0.156	-0.223	0.38***
		Twins			CSS	
	Boys	Girls	Δ	Boys	Girls	Δ
Score	n=19, 232	n=20, 209		n=5, 691	n=5, 834	
Total	0.039	-0.068	0.11***	-0.188	-0.245	0.06***
Reading	-0.075	0.062	0.14***	-0.353	-0.137	-0.22***
Math	0.181	-0.200	0.38***	0.072	-0.301	-0.37***

Notes: The complete sample entails all children for whom a test score is observed and gender can be identified between 2006 and 2014. The reported sample is discussed in more detail in section three. Test scores are standardized with mean zero and standard deviation one.

Boys (on average) outperform girls in the math-domain, whereas girls outperform boys in the reading-domain (see e.g. Guiso et al. (2008); Fryer and Levitt (2010); OECD (2015)). The gender performance gaps observed for the studied sample confirm the pattern found in the literature (Table 4). Boys perform significantly better at math, and girls perform significantly better at reading. Gender differences in educational performance are visible for the full sample, but also the sub-samples of twins and closely spaced singletons. They are slightly more pronounced in twins.

5 Results

The results are shown in Table 5 (aggregate score) and Table 6 (reading and math). Five specifications are shown which differ in the inclusion of controls. Controls for the earnings capacity of the household are available for a limited sample only. The base specifications are estimated for this smaller sample in specification three and four.

Table 5 shows the results for the (standardized) aggregate score as outcome variable. The twin coefficient is positive and significant in specifications without controls, and becomes smaller and insignificant when controls are added. This clearly shows that twins and CSS are born into different families. The female indicator shows that girls perform significantly lower on this aggregate scores than boys (by approximately 7% of a standard deviation).

Table 5: Pooled estimation results (part I)

	Aggregate Cito-score (scale: 501-550)				
	(1)	(2)	(3)	(4)	(5)
Twin	0.226***	-0.00620	0.220***	-0.00180	-0.00665
	(0.025)	(0.021)	(0.026)	(0.023)	(0.023)
OS	0.000642	0.0104	-0.00605	0.00411	0.00531
	(0.030)	(0.025)	(0.032)	(0.027)	(0.027)
Female	-0.0641**	-0.0677***	-0.0651*	-0.0687**	-0.0672**
	(0.032)	(0.026)	(0.034)	(0.028)	(0.028)
Twin*Female	-0.0285	-0.0406	-0.0394	-0.0398	-0.0390
	(0.035)	(0.029)	(0.038)	(0.031)	(0.031)
OS*Female	0.0139	-0.0271	0.0153	-0.0168	-0.0177
	(0.039)	(0.033)	(0.042)	(0.036)	(0.036)
Twin*OS	0.00397	-0.0404	-0.0198	-0.0466	-0.0467
	(0.034)	(0.028)	(0.036)	(0.031)	(0.030)
Twin*OS*Female	-0.0515	-0.00416	-0.0336	-0.00517	-0.00471
	(0.044)	(0.037)	(0.047)	(0.040)	(0.040)
D_{male}	-0.00397	0.0404	0.0198	0.0466	0.0467
	(0.034)	(0.028)	(0.036)	(0.031)	(0.030)
D_{female}*	-0.0476	-0.0445	-0.0534	-0.0517*	-0.0514*
	(0.034)	(0.028)	(0.036)	(0.030)	(0.030)
N	50,966	50,966	43,069	43,069	43,069
Controls	No	Yes	No	Yes	Yes
Income controls	No	No	No	No	Yes

[*] Adds the coefficients of $Twin{*}OS$ and $Twin{*}OS{*}Female$, equivalent to $\beta_6+\beta_7$.

[1] Estimated using OLS. The set of controls includes age, age squared, family size, birth order dummies, maternal age at birth, a non-native indicator, test-year dummies, household type dummies, indicator of whether the mother was in DI in the year of giving birth, and a control for the mean Cito-score at the school the child is attending in a given year. The additional household income controls contain a control for household income in the year the child turns four, and an indicator that the mother is working in this same year.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

Table 6: Pooled estimation results (part II)

	Reading score					Math score				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Twin	0.273***	0.0256	0.267***	0.0274	0.0222	0.112***	-0.0422**	0.114***	-0.0328	-0.0361
	(0.026)	(0.022)	(0.028)	(0.024)	(0.023)	(0.023)	(0.021)	(0.025)	(0.023)	(0.023)
OS	-0.00488	0.00320	-0.00380	0.00488	0.00616	-0.00297	0.00575	-0.0157	-0.00644	-0.00562
	(0.031)	(0.026)	(0.033)	(0.028)	(0.028)	(0.028)	(0.025)	(0.030)	(0.027)	(0.027)
Female	0.204***	0.203***	0.207***	0.208***	0.210***	-0.378***	-0.388***	-0.382***	-0.393***	-0.392***
	(0.032)	(0.027)	(0.035)	(0.029)	(0.029)	(0.030)	(0.027)	(0.033)	(0.029)	(0.029)
Twin*Female	-0.0608*	-0.0740**	-0.0736*	-0.0777**	-0.0769**	0.0145	0.00886	0.00878	0.0136	0.0142
	(0.036)	(0.030)	(0.039)	(0.032)	(0.032)	(0.034)	(0.030)	(0.037)	(0.032)	(0.032)
OS*Female	0.0243	-0.0157	0.0151	-0.0188	-0.0197	0.00819	-0.0203	0.0237	0.00280	0.00196
	(0.040)	(0.034)	(0.043)	(0.037)	(0.037)	(0.038)	(0.034)	(0.041)	(0.037)	(0.037)
Twin*OS	0.0115	-0.0304	-0.0141	-0.0399	-0.0400	-0.00808	-0.0424	-0.0262	-0.0462	-0.0464
	(0.035)	(0.029)	(0.038)	(0.032)	(0.032)	(0.032)	(0.028)	(0.034)	(0.031)	(0.031)
Twin*OS*Female	-0.0395	0.00723	-0.0156	0.0147	0.0151	-0.0549	-0.0191	-0.0487	-0.0294	-0.0289
	(0.044)	(0.038)	(0.048)	(0.042)	(0.042)	(0.043)	(0.039)	(0.046)	(0.042)	(0.042)
D_{male}	-0.0115	0.0304	0.0141	0.0399	0.0400	0.00808	0.0424	0.0262	0.0462	0.0464
	(0.035)	(0.029)	(0.038)	(0.032)	(0.032)	(0.032)	(0.028)	(0.034)	(0.031)	(0.031)
$D_{female}{}^*$	-0.0280	-0.0232	-0.0297	-0.0252	-0.0249	-0.0630*	-0.0615**	-0.0749**	-0.0755**	-0.0752**
	(0.033)	(0.028)	(0.036)	(0.031)	(0.031)	(0.033)	(0.029)	(0.036)	(0.032)	(0.032)
N	50,966	50,966	43,069	43,069	43,069	50,966	50,966	43,069	43,069	43,069
Controls	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes
Income controls	No	No	No	No	Yes	No	No	No	No	Yes

[*] Adds the coefficients of $Twin * OS$ and $Twin * OS * Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

The effects of twin testosterone transfers for boys (D_{male}) and for girls (D_{female}) are not significantly different from zero. If anything, the effect for females is negative and females with a male uterus-mate would perform about 5% of a standard deviation lower on the aggregate score, when controlling for the socialization effect of growing up with a brother.

The findings for the reading and math sub-score are shown in respectively the left and right panel of Table 6. The twin coefficient shows that twins are different from CSS and that including controls removes these differences. Women have a significant advantage in reading (by 2% of a standard deviation), whereas boys have an advantage in the math-domain (by about 4% of a standard deviation). The coefficients of interest: D_{male} and D_{female} are not significant for reading. However, girls with a male twin perform significantly worse on math by 7% of a standard deviation, even after controlling for socialization.[11]

This finding can be considered counterintuitive. One would expect that if boys have more prenatal testosterone than girls, and if girls with a twin brother are exposed to higher concentrations of prenatal testosterone, girls with a twin brother would show more male-typical performance patterns. Extrapolating this would lead to improved math performance and worse reading performance. We do not find support for the hypothesis that prenatal testosterone improves math performance or worsens language performance for girls. We also do not find improved math scores or lower language scores for boys. Hence we do not find evidence that prenatal testosterone amplifies (average) gender-specific differences.

Previous research also shows gender differences in test-score distributions (Halpern et al., 2007; Machin and Pekkarinen, 2008). Table A3, A4, and A5 show the results for re-estimating the models with indicators for scoring in the bottom 10%, bottom 25%, top 50%, top 25%, and top 10% in the three test-scores as outcome variables.[12] Girls exposed to higher levels of prenatal testosterone are 3.7% and 3.1% less likely to score in the top 50% for respectively the aggregate and math score. Males exposed to higher concentrations of prenatal testosterone are 1.9% less likely to score in the bottom 10% for both the aggregate and math score. Additionally, girls exposed to higher prenatal testosterone concentrations are 2.5% more likely to score in the bottom 10% of the math test-score distribution. The results on girls' average math performance might be driven by more girls scoring in the bottom 10%.

[11] Table A1 and Table A2 show that the results are robust to estimating the models separately for boys and girls.

[12] Coefficients are not different from the OLS estimates when estimated with quantile regression (full set of controls) as shown in Figure A1, A2, and A3.

5.1 Robustness checks

5.1.1 Different groups

A potential concern for our identification might be that maternal levels of testosterone are lower if spacing between children is less than four years (Maccoby et al., 1979; Baron-Cohen et al., 2004). We address this issue by restricting the sample to first born children. The results are shown in Table 7 and 8, the baseline model specification includes all control variables. The coefficient estimates are similar, especially the double difference estimate for females. The latter effect is not significant in these specification, which is likely due to less precision because the number of observations halved.

CSS have birth dates which are at most 12 months apart, therefore interactions are expected to be more twin-like than for regular siblings. We extend this difference to eighteen months (Table 7 and 8). An advantage is the increase in observations which increases the precision of estimates, a disadvantage is that interactions between these siblings are less twin-like, which makes them a less suitable control group.

Table 7: Robustness to using different groups (part I)

	Aggregate score (std)		
	(1)	(2)	(3)
	Baseline	First born	CSS 18 months
Twin	-0.00665	0.0159	-0.0722***
	(0.023)	(0.036)	(0.012)
OS	0.00531	-0.0158	0.0189**
	(0.027)	(0.046)	(0.008)
Female	-0.0672**	-0.0422	-0.0483***
	(0.028)	(0.045)	(0.008)
Twin*Female	-0.0390	-0.0541	-0.0598***
	(0.031)	(0.049)	(0.016)
OS*Female	-0.0177	-0.00120	-0.0363***
	(0.036)	(0.063)	(0.011)
Twin*OS	-0.0467	-0.0330	-0.0628***
	(0.030)	(0.050)	(0.017)
Twin*OS*Female	-0.00471	-0.0246	0.0159
	(0.040)	(0.068)	(0.022)
D_{males}	0.0467	0.0330	0.0628***
	(0.030)	(0.050)	(0.022)
$D_{females}$	-0.0514*	-0.0578	-0.0469***
	(0.030)	(0.048)	(0.016)
N	43,069	19,576	132,650
Controls	Y	Y	Y
Income controls	Y	Y	Y

[*] Adds the coefficients of $Twin*OS$ and $Twin*OS*Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

Table 8: Robustness to using different groups (part II)

	Language score (std)			Math score (std)		
	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	First born	CSS 18 months	Baseline	First born	CSS 18 months
Twin	0.0222	0.0466	-0.0490	-0.0361	-0.0104	-0.0806***
	(0.023)	(0.037)	(0.012)	(0.023)	(0.037)	(0.012)
OS	0.00616	-0.0537	0.0331***	-0.00562	-0.00187	-0.00884
	(0.028)	(0.044)	(0.008)	(0.027)	(0.046)	(0.008)
Female	0.210***	0.211***	0.201***	-0.392***	-0.332***	-0.0344**
	(0.029)	(0.046)	(0.009)	(0.029)	(0.049)	(0.017)
Twin*Female	-0.0769**	-0.0751	-0.0700***	0.0142	-0.0337	-0.0101
	(0.032)	(0.050)	(0.017)	(0.032)	(0.053)	(0.011)
OS*Female	-0.0197	0.00170	-0.0480***	0.00196	-0.00702	-0.0101
	(0.037)	(0.065)	(0.011)	(0.037)	(0.066)	(0.011)
Twin*OS	-0.0400	-0.0222	-0.0692***	-0.0464	-0.0506	-0.0457***
	(0.032)	(0.052)	(0.017)	(0.031)	(0.050)	(0.017)
Twin*OS*Female	0.0151	0.0139	0.0453**	-0.0289	-0.0164	-0.0153
	(0.042)	(0.070)	(0.022)	(0.042)	(0.072)	(0.023)
D_{male}	0.0400	0.0222	0.692***	0.0464	0.0506	0.0457***
	(0.032)	(0.052)	(0.017)	(0.031)	(0.050)	(0.017)
D_{female}*	-0.0249	-0.0361	-0.0240	-0.0752**	-0.0671	-0.0610***
	(0.031)	(0.048)		(0.032)	(0.052)	(0.018)
N	43,069	19,576	132,650	43,069	19,576	132,650
Controls	Y	Y	Y	Y	Y	Y
Income controls	Y	Y	Y	Y	Y	Y

* Adds the coefficients of $Twin * OS$ and $Twin * OS * Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

The double difference estimates for females are very robust to using different bandwidths. The effects the aggregate score are approximately -5% of a standard deviation, for language they are -2% of a standard deviation, and for math they are between -6% and -7% of a standard deviation. The larger the difference the more significant findings are which is due to the larger number of observations that increases the precision of the estimates. The double difference estimates for males are less robust. For math they are quite constant at 5% of a standard deviation, whereas they range between 5% to 7%, and 4% and 7% of a standard deviation for the aggregate and language score respectively.

The opposite-sex coefficient increases, implying that gender-mixed socialization for boys is larger when sibling spacing increases. However, this coefficient drops out when we look at the double difference estimator, which is simply the inverse of the $Twin * OS$ coefficient. This coefficient becomes larger (more negative), which explains the larger double difference estimator. Hence boys who grow up with sister perform worse, and boys who who have a twin brother do better. As the estimates increase when the control group is less conservatively chosen point-estimates for the effect for boys increase. Which likely implies that our baseline estimate is conservative.

5.1.2 Matching estimators

We know that twins and CSS are born into different families and hence we control for many of these differences. This section will employ matching estimators to make the sample of CSS and twins more comparable before estimating the parameters (Table A6).

We employ Kernel matching (Epanechnikov kernel with a bandwidth of 0.06), and weights to the observations are assigned with the Kernel matching procedure (column 1, 4 and 7). Inverse Probability Matching (IPM) is also used (column 2, 5, and 8), but as this method is very sensitive to very high and low propensity scores a more robust type will be used that only includes observations with propensity score between 0.1 and 0.9 (column 3, 6, and 9).

Table A6 shows that IPM is very sensitive to excluding those with very high and low propensity scores, and hence it is better to only look at those results that exclude these observations. The estimates using Kernel matching are all larger than the baseline estimates, which implies that our specification with controls gives a conservative estimate of the true effect. The IPM specification gives smaller double difference estimates for boys and larger double difference estimates for girls. However the estimates still confirm that there is no effect for boys. Whereas the effect for girls would be larger with matching. Hence the estimates for girls are conservative in the main specification with controls.

6 Discussion

The result that girls that are exposed to higher concentrations of prenatal testosterone perform 7% of a standard deviation lower on math is counterintuitive. One would expect that these girls would me more male-typical and hence their educational performance would also be more male-typical. Gielen et al. (2016) do not find increased earnings for for females exposed to higher prenatal testosterone. This is consistent with our finding, especially since math performance is related to earnings are related (e.g. **?**). They explain their null-finding, and if anything negative effect, with labor market discrimination for masculinized females. This section provides several explanations for the negative effect found for females on educational performance at age twelve.

We are interested in the effect of prenatal testosterone (T) on educational outcomes (Y). However we cannot rule out that besides there being a direct effect of prenatal testosterone on educational performance (first term), testosterone might be interacting with external factors (E), that can in turn affect educational outcomes (second term).

$$\frac{dY}{dT} = \underbrace{\frac{\delta Y}{\delta T}}_{1} + \underbrace{\frac{\delta Y}{\delta E}\frac{\delta E}{\delta T}}_{2} \tag{2}$$

It could be that prenatal testosterone directly shapes mathematics performance by affecting brain development (as suggested by Jordan-Young (2010)), hence the second term would be equal to zero. This would imply that prenatal testosterone worsens mathematics performance for girls. Boys might not experience such negative effects because of later life factors (e.g. toys they play with, gender stereotypes that shape preferences), which enhances their math performance later in life. Potentially it could also be that a little extra prenatal testosterone does not affect boys much, as they are already exposed to high concentrations (Resnick et al., 1993; Peper et al., 2009; Tapp et al., 2011; Cronqvist et al., 2015).

We cannot rule out that biological factors, like prenatal testosterone exposure, and social factors (i.e. culture) interact with one another as has been found before (Kendler et al., 1995; Cadoret et al., 1996; Turkheimer et al., 2003; Sacerdote, 2007). The latter indicates that environmental impacts are larger when a child is exposed to a poor socioeconomic background. Hence prenatal testosterone might express itself differently when children grow up in different environments.

A different argument could be that opposite-sex twinning is related to other birth outcomes that might affect educational performance. Boys on average weigh more than girls at birth. As a result one could expect that sharing the intrauterine environment with an opposite-sex fetus might affect birth weight. Birth weight in mice is higher for females located between two male fetuses as opposed to females located between two female fetuses (Miller and Martin, 1995). The evidence for humans in mixed. Females from opposite twin pairs have higher birth weights than females from same-sex twin pairs (Glinianaia et al., 1998; Blickstein and Kalish, 2003). This relationship is not confirmed by Orlebeke et al. (1993). Loos et al. (2001) find instead that boys from opposite-sex twin pairs weigh more than boys from same-sex pairs, whereas no such relationship is found for females. Hence there is no conclusive evidence that birth weight is affected by the sex of the co-twin.

Another explanation could be that girls with more prenatal testosterone are more male-typical in morphological characteristics and behavior (e.g. Cohen-Bendahan et al. (2004); Peper et al. (2009); Vuoksimaa et al. (2010a), and this might eventually affect educational performance through external factors. Perhaps girls exposed to higher concentrations are more male-typical, and therefore insecure, which makes them perform worse. Or perhaps they look and behave more male-typical and try to oppose this maleness by confirming female gender stereotypes.[13] It could also work through the parental channel. Yi et al. (2015) study health shocks in twin pairs and find that extra health investments for the twin who experiences the health shock are compensated for by less educational investments. This could fit our story if the girl who was exposed to higher prenatal testosterone concentrations is more masculine, which requires other investments by parents, which comes at the cost of educational investments of parents.

We consider three different educational outcomes, which might help in explaining the negative effect found for females. First school advice is examined, which can be regarded (partly) as a teacher assessment of the child's ability. School advice is hierarchical with one being the lowest and nine being the highest. It is based upon teacher assessments of the child's ability and the child's performance on standardized tests over the course of his or her primary school career. This outcome variable allows us to study whether teachers give different school advice if children differ w.r.t. their prenatal testosterone levels. It is an imperfect measure as standardized tests and overall test performance also play a role in school advice. The variable is not available for the full sample, but results (Table 11) show that the double difference estimates for females and males are not significantly different from zero. The signs are consistent with the main results, if anything females exposed to higher prenatal testosterone concentrations receive a lower school advice, and the effect is opposite for men.

The second outcome regards the child's score on an optional part of the test: world orientation. Schools can choose to participate in this part,

[13] Similar to "acting White" where Black individuals are punished by peers for acting differently, and as a consequence are not incentivized to act in particular "White" ways (see e.g. **?**).

but it does not count towards the final score, implying that it might serve as a proxy for motivation. Children know that this part does not count towards the final score, but they have to complete it. The measure is imperfect as it might also capture actual ability for this specific task. The double difference estimates are insignificant for boys and girls.

Table 9: Other educational outcomes

	Teacher assessment (scale=1-9)	Optional part of test (standardized)	Scoring above school average average (0-1)
Twin	0.00856	0.00957	-0.00659***
	(0.0746)	(0.026)	(0.012)
OS	-0.0298	0.0312	0.000594
	(0.0879)	(0.032)	(0.015)
Female	-0.128	-0.0365***	-0.0515***
	(0.0919)	(0.033)	(0.015)
Twin*Female	-0.0137	-0.0850**	-0.00401
	(0.104)	(0.037)	(0.017)
OS*Female	-0.0351	-0.0440	0.0106
	(0.119)	(0.043)	(0.020)
Twin*OS	-0.101	-0.0452	-0.0188
	(0.102)	(0.036)	(0.017)
Twin*OS*Female	0.0314	0.0156	-0.0200
	(0.135)	(0.048)	(0.022)
D_{males}	0.101	0.0452	0.0188
	(0.102)	(0.036)	(0.017)
$D_{females}$	-0.0693	-0.0296	-0.0388**
	(0.102)	(0.036)	(0.016)
N	30,944	41,527	43,096
Controls	Y	Y	Y
Income controls	Y	Y	Y

[*] Adds the coefficients of $Twin*OS$ and $Twin*OS*Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

The third outcome variable is an indicator for whether the child scored above the school average in a given year. This can be seen as a proxy for competitiveness, although it is imperfect as it also captures actual performance. The double difference estimator for girls shows that they are about 4% less likely to score above school average in a given year when they had a male uterus mate. For boys such a significance difference is not detectable. Once again, this tendency for girls to score below average could be caused by the fact that this group scores significantly lower on math.

7 Conclusion

On average boys perform better at math and girls perform better at reading. Little is known about the role of biology in creating these gender differences. This paper examines the role of biology and specifically the role of prenatal testosterone. Prenatal testosterone is responsible for the sexual differentiation of the male fetus and is also said to affect brain development. Twinning is used as an exogenous proxy for prenatal testosterone as it is impossible to directly relate prenatal testosterone and educational outcomes.

Males are exposed to elevated levels of prenatal testosterone between the eighth and twenty-fourth week of gestation. This testosterone can transfer from the male twin to his uterus mate. Individuals with a male co-twin are exposed to higher levels of prenatal testosterone than individuals with a female co-twin. Females with a male co-twin are more masculine in morphological characteristics (e.g. more masculine 2D:4D ratio), behavior, and cognitive capacities. Whereas for males usually no increased masculine characteristics or behavior is found.

A control group of closely spaced singletons (CSS) is used to isolate the effect of prenatal testosterone. CSS are siblings whose birth dates are at most twelve months apart. The effect of prenatal testosterone can be isolated if socialization is similar for this group and twins. The twelve month window is small and hence it is likely that siblings born very near each other experience similar environments, and that their interactions are more twin-like than for siblings born outside the 12-month window.

We use administrative data from Statistics Netherlands with information on a standardized test performed in the final year of primary education and find that prenatal testosterone does not alter educational performance for males. For females no effects are found on an aggregate score and a reading score. When controlling for socialization, girls with a male co-twin, who are exposed to higher levels of prenatal testosterone, are performing 7% of a standard deviation lower on math. This effect can be explained by the fact that more women with a male twin end up in the lowest 10% of the test-score distribution.

The latter finding is counterintuitive as one would expect improved performance for girls who are more male-typical due to higher concentrations of prenatal testosterone in utero. Possible explanations could be that prenatal testosterone actually causes lower math performance, and that boys make up for this disadvantage in their youth (e.g. by playing with different toys). Another explanation could be that girls with a male co-twin are more masculine, and that this masculinity affects educational performance negatively at age twelve.

We do not find evidence that prenatal testosterone shapes gender differentials as we observe them: it does not improve math performance, or worsen language performance. Hence we do not find evidence for a role of biology, and specifically prenatal testosterone, in determining educational gender differentials. As research on the role of biology in determining gender differentials in any domain is limited, future research should address whether other biological factors play a role in determining these gender differentials.

References

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

Almond, D. and Currie, J. (2011). Killing me softly: The fetal origins hypothesis. *The Journal of Economic Perspectives*, 25(3):153–172.

Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, pages 450–477.

Austin, E. J., Manning, J. T., McInroy, K., and Mathews, E. (2002). A preliminary investigation of the associations between personality, cognitive ability and digit ratio. *Personality and individual differences*, 33(7):1115–1124.

Auyeung, B., Lombardo, M. V., and Baron-Cohen, S. (2013). Prenatal and postnatal hormone effects on the human brain and cognition. *Pflügers Archiv-European Journal of Physiology*, 465(5):557–571.

Banda, I., Tagne, A., Chew, H., Vigneswara Ilavarasan, P., Levy, M., Gilbert, M. R., Masucci, M., Gilbert, M. R., Masucci, M., Klonner, S., et al. (2010). *World development report 2012: gender equality and development.* The International Bank for Reconstruction and Development/The World Bank.

Baron-Cohen, S., Lutchmaya, S., and Knickmeyer, R. (2004). *Prenatal testosterone in mind: Amniotic fluid studies.* MIT Press.

Bettencourt, B. and Miller, N. (1996). Gender differences in aggression as a function of provocation: a meta-analysis. *Psychological bulletin*, 119(3):422.

Bhalotra, S. R., Clarke, D., et al. (2016). The twin instrument. Technical report, Institute for the Study of Labor (IZA).

Bharadwaj, P., De Giorgi, G., Hansen, D. R., and Neilson, C. (2015). The gender gap in mathematics: evidence from a middle-income country. *FRB of New York Working Paper No. FEDNSR721.*

Blau, F. D. and Kahn, L. M. (2000). Gender differences in pay. Technical report, National bureau of economic research.

Blau, F. D. and Kahn, L. M. (2016). The gender wage gap: Extent, trends, and explanations. *IZA Discussion Paper.*

Blickstein, I. and Kalish, R. B. (2003). Birthweight discordance in multiple pregnancy. *Twin Research*, 6(06):526–531.

Bronars, S. G. and Grogger, J. (1994). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *The American Economic Review*, pages 1141–1156.

Buser, T. (2012a). Digit ratios, the menstrual cycle and social preferences. *Games and Economic Behavior*, 76(2):457–470.

Buser, T. (2012b). The impact of the menstrual cycle and hormonal contraceptives on competitiveness. *Journal of Economic Behavior & Organization*, 83(1):1–10.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, 129(3):1409–1447.

Cadoret, R. J., Winokur, G., Langbehn, D., Troughton, E., et al. (1996). Depression spectrum disease, i: The role of gene-environment interaction. *The American journal of psychiatry*, 153(7):892.

Carneiro, P. M. and Heckman, J. J. (2003). Human capital policy.

Ceci, S. J., Williams, W. M., and Barnett, S. M. (2009). Women's underrepresentation in science: sociocultural and biological considerations. *Psychological bulletin*, 135(2):218.

Coates, J. M., Gurnellc, M., and Rustichinid, A. (2009). Second-to-fourth digit ratio predicts success among high-frequency financial traders. *PNAS*, 106(2):623–628.

Cohen-Bendahan, C. C., Buitelaar, J. K., van Goozen, S. H., and Cohen-Kettenis, P. T. (2004). Prenatal exposure to testosterone and functional cerebral lateralization: a study in same-sex and opposite-sex twin girls. *Psychoneuroendocrinology*, 29(7):911–916.

Cohen-Bendahan, C. C., van de Beek, C., and Berenbaum, S. A. (2005). Prenatal sex hormone effects on child and adult sex-typed behavior: methods and findings. *Neuroscience & Biobehavioral Reviews*, 29(2):353–384.

Coolican, J. and Peters, M. (2003). Sexual dimorphism in the 2d/4d ratio and its relation to mental rotation performance. *Evolution and Human Behavior*, 24(3):179–183.

Cronqvist, H., Previtero, A., Siegel, S., and White, R. E. (2015). The fetal origins hypothesis in finance: Prenatal environment, the gender gap, and investor behavior. *Review of Financial Studies*, page hhv065.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature*, pages 448–474.

Currie, J. and Almond, D. (2011). Human capital development before age five. *Handbook of labor economics*, 4:1315–1486.

Derks, E. M., Dolan, C. V., and Boomsma, D. I. (2006). A test of the equal environment assumption (eea) in multivariate twin studies. *Twin Research and Human Genetics*, 9(3):403–411.

Dreber, A. and Hoffman, M. (2007). Portfolio selection in utero. *Stockholm School of Economics.*

Eckel, C. C. and Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, 1:1061–1073.

Ellison, G. and Swanson, A. (2009). The gender gap in secondary school mathematics at high achievement levels: Evidence from the american mathematics competitions. Technical report, National Bureau of Economic Research.

Eriksson, M., Rasmussen, F., and Tynelius, P. (2006). Genetic factors in physical activity and the equal environment assumption–the swedish young male twins study. *Behavior genetics*, 36(2):238–247.

Fellman, J. and Eriksson, A. W. (2006). Weinberg's differential rule reconsidered. *Human Biology*, 78(3):253–275.

Fellman, J. and Eriksson, A. W. (2010). Secondary sex ratio in multiple births. *Twin Research and Human Genetics*, 13(01):101–108.

Flory, J. A., Leibbrandt, A., and List, J. A. (2010). Do competitive work places deter female workers? a large-scale natural field experiment on gender differences in job-entry decisions. Technical report, National Bureau of Economic Research.

Fryer, R. G. and Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, 2(2):210–240.

Garbarino, E., Slonim, R., and Sydnor, J. (2011). Digit ratios (2d: 4d) as predictors of risky decision making for both sexes. *Journal of Risk and Uncertainty*, 42(1):1–26.

Gielen, A. C., Holmes, J., and Myers, C. (2016). Prenatal testosterone and the earnings of men and women. *Journal of Human Resources*, 51(1):30–61.

Glinianaia, S. V., Magnus, P., Harris, J. R., and Tambs, K. (1998). Is there a consequence for fetal growth of having an unlike-sexed cohabitant in utero? *International Journal of Epidemiology*, 27(4):657–659.

Gneezy, U., Niederle, M., Rustichini, A., et al. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *The American Economic Review*, 94(2):377–381.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science*, 320(5880):1164–1165.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., and Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological science in the public interest*, 8(1):1–51.

Hampson, E., Ellis, C. L., and Tenk, C. M. (2008). On the relation between 2d: 4d and sex-dimorphic personality traits. *Archives of sexual behavior*, 37(1):133–144.

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic inquiry*, 46(3):289–324.

Heil, M., Kavšek, M., Rolke, B., Beste, C., and Jansen, P. (2011). Mental rotation in female fraternal twins: Evidence for intra-uterine hormone transfer? *Biological psychology*, 86(1):90–93.

Hettema, J. M., Neale, M. C., and Kendler, K. S. (1995). Physical similarity and the equal-environment assumption in twin studies of psychiatric disorders. *Behavior genetics*, 25(4):327–335.

Hönekopp, J., Bartholdt, L., Beier, L., and Liebert, A. (2007). Second to fourth digit length ratio (2d: 4d) and adult sex hormone levels: new data and a meta-analytic review. *Psychoneuroendocrinology*, 32(4):313–321.

Jacobsen, J. P., Pearce III, J. W., and Rosenbloom, J. L. (1999). The effects of child-bearing on married women's labor supply and earnings: using twin births as a natural experiment. *Journal of Human Resources*, pages 449–474.

James, W. H. (2010). The sex ratios of monozygotic and dizygotic twins. *Twin research and human genetics*, 13(4):381–382.

Jordan-Young, R. M. (2010). *Brain storm.* Harvard University Press.

Kendler, K. S., Kessler, R. C., Walters, E. E., MacLean, C., Neale, M. C., Heath, A. C., and Eaves, L. J. (1995). Stressful life events, genetic liability, and onset of an episode of major depression in women. *Am J Psychiatry*, 152:842.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., and Eaves, L. J. (1994). Parental treatment and the equal environment assumption in twin studies of psychiatric illness. *Psychological medicine*, 24(3):579–590.

Knudsen, E. I., Heckman, J. J., Cameron, J. L., and Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building americas future workforce. *Proceedings of the National Academy of Sciences*, 103(27):10155–10162.

Loos, R. J., Derom, C., Eeckels, R., Derom, R., and Vlietinck, R. (2001). Length of gestation and birthweight in dizygotic twins. *The Lancet*, 358(9281):560–561.

LoParo, D. and Waldman, I. (2014). Twins rearing environment similarity and childhood externalizing disorders: A test of the equal environments assumption. *Behavior genetics*, 44(6):606–613.

Lutchmaya, S., Baron-Cohen, S., Raggatt, P., Knickmeyer, R., and Manning, J. T. (2004). 2nd to 4th digit ratios, fetal testosterone and estradiol. *Early human development*, 77(1):23–28.

Maccoby, E. E., Doering, C. H., Jacklin, C. N., and Kraemer, H. (1979). Concentrations of sex hormones in umbilical-cord blood: their relation to sex and birth order of infants. *Child development*, pages 632–642.

Machin, S. and Pekkarinen, T. (2008). Global sex differences in test score variability. *Science.*

Manning, J. T. and Taylor, R. P. (2001). Second to fourth digit ratio and male ability in sport: implications for sexual selection in humans. *Evolution and Human Behavior*, 22(1):61–69.

Matheny, A. P., Wilson, R. S., and Dolan, A. B. (1976). Relations between twins' similarity of appearance and behavioral similarity: Testing an assumption. *Behavior Genetics*, 6(3):343–351.

Medland, S. E., Loehlin, J. C., and Martin, N. G. (2008). No effects of prenatal hormone transfer on digit ratio in a large sample of same-and opposite-sex dizygotic twins. *Personality and Individual Differences*, 44(5):1225–1234.

Miller, D. I. and Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in cognitive sciences*, 18(1):37–45.

Miller, E. M. (1994). Prenatal sex hormone transfer: A reason to study opposite-sex twins. *Personality and Individual Differences*, 17(4):511–529.

Miller, E. M. and Martin, N. (1995). Analysis of the effect of hormones on opposite-sex twin attitudes. *Acta geneticae medicae et gemellologiae: twin research*, 44(01):41–52.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, pages 1067–1101.

Niederle, M. and Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *The Journal of Economic Perspectives*, 24(2):129–144.

Nollenberger, N., Rodriguez Planas, N., and Sevilla Sanz, A. (2014). The math gender gap: The role of culture. *IZA Discussion Paper.*

Nye, J. and Orel, E. (2015). The influence of prenatal hormones on occupational choice: 2d: 4d evidence from moscow. *Personality and Individual Differences*, 78:39–42.

OECD (2015). The abc of gender equality in education: Aptitude, behaviour, confidence. Technical report, OECD Publishing.

Orlebeke, J. F., Caroline, G., van Baal, M., Boomsma, D. I., and Neeleman, D. (1993). Birth weight in opposite sex twins as compared to same sex dizygotic twins. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 50(2):95–98.

Örs, E., Palomino, F., and Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, 31(3):443–499.

Peper, J. S., Brouwer, R. M., Van Baal, G. C. M., Schnack, H. G., Van Leeuwen, M., Boomsma, D. I., Kahn, R. S., and Pol, H. E. H. (2009). Does having a twin brother make for a bigger brain? *European Journal of Endocrinology*, 160(5):739–746.

Pope, D. G. and Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *The Journal of Economic Perspectives*, 24(2):95–108.

Puts, D. A., McDaniel, M. A., Jordan, C. L., and Breedlove, S. M. (2008). Spatial ability and prenatal androgens: Meta-analyses of congenital adrenal hyperplasia and digit ratio (2d: 4d) studies. *Archives of sexual behavior*, 37(1):100–111.

Resnick, S. M., Gottesman, I. I., and McGue, M. (1993). Sensation seeking in opposite-sex twins: an effect of prenatal hormones? *Behavior Genetics*, 23(4):323–329.

Rosenzweig, M. R. and Wolpin, K. I. (1980a). Life-cycle labor supply and fertility: Causal inferences from household models. *The Journal of Political Economy*, pages 328–348.

Rosenzweig, M. R. and Wolpin, K. I. (1980b). Testing the quantity-quality fertility model: The use of twins as a natural experiment. *Econometrica: journal of the Econometric Society*, pages 227–240.

Sacerdote, B. (2007). How large are the effects from changes in family environment? a study of korean american adoptees. *The Quarterly Journal of Economics*, 122(1):119–157.

Scarr, S. and Carter-Saltzman, L. (1979). Twin method: Defense of a critical assumption. *Behavior genetics*, 9(6):527–542.

Scholte, R. S., Van den Berg, G. J., and Lindeboom, M. (2015). Long-run effects of gestation during the dutch hunger winter famine on labor market and hospitalization outcomes. *Journal of health economics*, 39:17–30.

Slutske, W. S., Bascom, E. N., Meier, M. H., Medland, S. E., and Martin, N. G. (2011). Sensation seeking in females from opposite-versus same-sex twin pairs: hormone transfer or sibling imitation? *Behavior genetics*, 41(4):533–542.

Speiser, P. W. and White, P. C. (2003). Congenital adrenal hyperplasia. *New England Journal of Medicine*, 349(8):776–788.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35(1):4–28.

Stoet, G. and Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, 16(1):93.

Stoet, G. and Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of pisa data. *PloS one*, 8(3):e57988.

Tapp, A. L., Maybery, M. T., and Whitehouse, A. J. (2011). Evaluating the twin testosterone transfer hypothesis: a review of the empirical evidence. *Hormones and behavior*, 60(5):713–722.

Turkheimer, E., Haley, A., Waldron, M., d'Onofrio, B., and Gottesman, I. I. (2003). Socioeconomic status modifies heritability of iq in young children. *Psychological science*, 14(6):623–628.

van Anders, S. M., Vernon, P. A., and Wilbur, C. J. (2006). Finger-length ratios show evidence of prenatal hormone-transfer between opposite-sex twins. *Hormones and Behavior*, 49(3):315–319.

Van de Beek, C., Thijssen, J. H., Cohen-Kettenis, P. T., van Goozen, S. H., and Buitelaar, J. K. (2004). Relationships between sex hormones assessed in amniotic fluid, and maternal and umbilical cord serum: what is the best source of information to investigate the effects of fetal hormonal exposure? *Hormones and Behavior*, 46(5):663–669.

Vlietinck, R., Derom, C., Derom, R., Van den Berghe, H., and Thiery, M. (1988). The validity of weinberg's rule in the east flanders prospective twin survey (efpts). *AMG Acta geneticae medicae et gemellologiae: twin research*, 37(2):137–141.

Voracek, M. and Dressler, S. G. (2007). Digit ratio (2d: 4d) in twins: heritability estimates and evidence for a masculinized trait expression in women from opposite-sex pairs. *Psychological reports*, 100(1):115–126.

Vuoksimaa, E., Eriksson, C. P., Pulkkinen, L., Rose, R. J., and Kaprio, J. (2010a). Decreased prevalence of left-handedness among females with male co-twins: evidence suggesting prenatal testosterone transfer in humans? *Psychoneuroendocrinology*, 35(10):1462–1472.

Vuoksimaa, E., Kaprio, J., Kremen, W. S., Hokkanen, L., Viken, R. J., Tuulio-Henriksson, A., and Rose, R. J. (2010b). Having a male co-twin masculinizes mental rotation performance in females. *Psychological science*.

Wilder, G. Z. and Powell, K. (1989). Sex differences in test performance: A survey of the literature. *ETS Research Report Series*, 1989(1):i–50.

Yi, J., Heckman, J. J., Zhang, J., and Conti, G. (2015). Early health shocks, intrahousehold resource allocation and child outcomes. *The Economic Journal*, 125(588).

Appendix

Table A1: Pooled estimation results (part I)

	Male twins and closely spaced singletons				
	Aggregate Cito-score (scale: 501-550)				
	(1)	(2)	(3)	(4)	(5)
Twin	0.226***	-0.00180	0.220***	0.00308	-0.000719
	(0.025)	(0.021)	(0.026)	(0.023)	(0.023)
OS	0.000642	0.0107	-0.00605	0.00423	0.00528
	(0.030)	(0.025)	(0.032)	(0.027)	(0.027)
Twin*OS	0.00397	-0.0404	-0.0198	-0.0467	-0.0468
	(0.034)	(0.028)	(0.036)	(0.031)	(0.030)
N	24,923	24,923	21,030	21,030	21,030
Controls	No	Yes	No	Yes	Yes
Income controls	No	No	No	No	Yes

	Female twins and closely spaced singletons				
	Aggregate Cito-score (scale: 501-550)				
	(1)	(2)	(3)	(4)	(5)
Twin	0.197***	-0.05068**	0.180***	-0.0455*	-0.0507**
	(0.025)	(0.022)	(0.027)	(0.023)	(0.023)
OS	0.0146	-0.0162	0.00927	-0.0108	-0.0106
	(0.030)	(0.025)	(0.032)	(0.027)	(0.027)
Twin*OS	-0.0476	-0.0464*	-0.0534	-0.0542*	-0.0539*
	(0.034)	(0.028)	(0.036)	(0.030)	(0.030)
N	26,043	26,043	22,039	22,039	22,039
Controls	No	Yes	No	Yes	Yes
Income controls	No	No	No	No	Yes

[*] Adds the coefficients of $Twin * OS$ and $Twin * OS * Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

Table A2: Pooled estimation results (part II)

	Male twins and closely spaced singletons									
	Language score					Math score				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Twin	0.273***	0.0208	0.267***	0.0223	0.0182	0.112***	-0.0294	0.114***	-0.0187	-0.0211
	(0.026)	(0.022)	(0.028)	(0.024)	(0.024)	(0.023)	(0.021)	(0.025)	(0.023)	(0.023)
OS	-0.00488	0.00321	-0.00380	0.00492	0.00606	-0.00297	0.00638	-0.0157	-0.00598	
	(0.031)	(0.026)	(0.033)	(0.028)	(0.028)	(0.028)	(0.024)	(0.030)	(0.027)	(0.027)
Twin*OS	0.0115	-0.0317	-0.0141	-0.0412	-0.0413	-0.00808	-0.0409	-0.0262	-0.0454	-0.0455
	(0.032)	(0.029)	(0.038)	(0.032)	(0.032)	(0.032)	(0.028)	(0.034)	(0.031)	(0.030)
N	24,923	24,923	21,030	21,030	21,030	24,923	24,923	21,030	21,030	21,030
Controls	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes
Income controls	No	No	No	No	Yes	No	No	No	No	Yes
	Female twins and closely spaced singletons									
	Language score					Math score				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Twin	0.212***	-0.0442**	0.194***	-0.0462*	-0.0516**	0.126***	-0.0446**	0.122***	-0.0305	-0.0342
	(0.025)	(0.022)	(0.027)	(0.024)	(0.024)	(0.025)	(0.023)	(0.027)	(0.025)	(0.025)
OS	0.0194	-0.0113	0.0113	-0.0116	-0.0112	0.00522	-0.0145	0.00793	-0.00213	-0.00215
	(0.029)	(0.025)	(0.032)	(0.027)	(0.027)	(0.029)	(0.026)	(0.032)	(0.028)	(0.028)
Twin*OS	-0.0280	-0.0242	-0.0297	-0.0268	-0.0266	-0.0630*	0.0643**	-0.0749**	-0.0789**	-0.0785**
	(0.033)	(0.028)	(0.036)	(0.031)	(0.030)	(0.033)	(0.030)	(0.036)	(0.032)	(0.032)
N	26,043	26,043	22,039	22,039	22,039	26,043	26,043	22,039	22,039	22,039
Controls	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes
Income controls	No	No	No	No	Yes	No	No	No	No	Yes

[*] Adds the coefficients of $Twin * OS$ and $Twin * OS * Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

Table A3: Alternative quantile regression

	Aggregate score (std)				
	(1)	(2)	(3)	(4)	(5)
	I[Bottom 10%]	I[Bottom 25%]	I[Top 50%]	I[Top 25%]	I[Top 10%]
Twin	0.00392	0.00968	-0.00284	-0.00456	0.00343
	(0.007)	(0.010)	(0.011)	(0.010)	(0.007)
OS	0.00453	0.00141	0.000943	0.00968	0.0106
	(0.009)	(0.013)	(0.014)	(0.012)	(0.008)
Female	0.0151	0.0357***	-0.0436***	-0.0222*	0.000684
	(0.010)	(0.013)	(0.014)	(0.012)	(0.008)
Twin*Female	0.000102	-0.000876	-0.00699	-0.0227*	-0.0269***
	(0.011)	(0.015)	(0.016)	(0.013)	(0.010)
OS*Female	0.000151	0.00618	0.00578	-0.00783	-0.0162
	(0.013)	(0.017)	(0.019)	(0.015)	(0.011)
Twin*OS	0.00623	0.0127	-0.0159	-0.0215	-0.0139
	(0.010)	(0.014)	(0.016)	(0.014)	(0.010)
Twin*OS*Female	0.00440	0.00767	-0.0206	0.00638	0.0136
	(0.015)	(0.020)	(0.021)	(0.018)	(0.013)
D_{males}	-0.0191**	-0.0199	0.000934	0.00180	0.00901
	(0.009)	(0.013)	(0.016)	(0.015)	(0.011)
$D_{females}$	0.0106	0.0204	-0.0365**	-0.0151	-0.000322
	(0.011)	(0.015)	(0.016)	(0.013)	(0.009)
N	43,069	43,069	43,069	43,069	43,069
Controls	Y	Y	Y	Y	Y
Income controls	Y	Y	Y	Y	Y

[*] Adds the coefficients of $Twin * OS$ and $Twin * OS * Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

Table A4: Alternative quantile regression

	Reading score (std)				
	(1)	(2)	(3)	(4)	(5)
	I[Bottom 10%]	I[Bottom 25%]	I[Top 50%]	I[Top 25%]	I[Top 10%]
Twin	-0.0170**	0.00587	0.000344	-0.00413	0.00406
	(0.008)	(0.011)	(0.011)	(0.009)	(0.006)
OS	0.00267	0.00796	0.00683	0.0127	0.00612
	(0.011)	(0.013)	(0.013)	(0.011)	(0.007)
Female	-0.0502***	-0.0630***	0.0800***	0.0515***	0.0310***
	(0.010)	(0.013)	(0.014)	(0.012)	(0.008)
Twin*Female	0.0226**	0.00982	-0.0267*	-0.0213	0.0176**
	(0.011)	(0.015)	(0.016)	(0.013)	(0.009)
OS*Female	-0.00597	0.00697	-0.0235	-0.00707	-0.00538
	(0.014)	(0.018)	(0.018)	(0.015)	(0.010)
Twin*OS	0.00672	0.00187	-0.0199	-0.0194	-0.0158*
	(0.012)	(0.015)	(0.015)	(0.013)	(0.008)
Twin*OS*Female	0.00746	-0.00134	0.0168	0.0100	0.0129
	(0.015)	(0.020)	(0.021)	(0.018)	(0.012)
D_{males}	-0.00672	-0.00187	0.0199	0.0194	0.0158*
	(0.012)	(0.015)	(0.015)	(0.013)	(0.008)
$D_{females}$	0.0142	0.00535	-0.00314	-0.00939	-0.00290
	(0.010)	(0.014)	(0.016)	(0.014)	(0.010)
N	43,069	43,069	43,069	43,069	43,069
Controls	Y	Y	Y	Y	Y
Income controls	Y	Y	Y	Y	Y

* Adds the coefficients of $Twin * OS$ and $Twin * OS * Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

Table A5: Alternative quantile regression

	Math score (std)				
	(1)	(2)	(3)	(4)	(5)
	I[Bottom 10%]	I[Bottom 25%]	I[Top 50%]	I[Top 25%]	I[Top 10%]
Twin	0.0116*	0.0165*	-0.130	-0.00994	0.0125
	(0.006)	(0.010)	(0.012)	(0.011)	(0.008)
OS	-0.00167	-0.00199	-0.0124	-0.00773	0.000632
	(0.008)	(0.012)	(0.014)	(0.013)	(0.009)
Female	0.0747***	0.132***	-0.177***	-0.120***	-0.0513***
	(0.009)	(0.013)	(0.014)	(0.012)	(0.008)
Twin*Female	-0.0152	-0.0115	0.0169	-0.0168	-0.0282***
	(0.010)	(0.014)	(0.016)	(0.014)	(0.010)
OS*Female	-0.00130	0.0131	0.0147	0.00367	0.00120
	(0.013)	(0.017)	(0.019)	(0.016)	(0.011)
Twin*OS	0.0191**	0.0199	-0.000934	-0.00180	-0.00901
	(0.009)	(0.013)	(0.016)	(0.015)	(0.011)
Twin*OS*Female	0.00601	-0.00402	-0.303	-0.0111	0.00347
	(0.014)	(0.020)	(0.022)	(0.019)	(0.013)
D_{males}	-0.0191**	-0.0199	0.000934	0.00180	0.00901
	(0.009)	(0.013)	(0.016)	(0.015)	(0.011)
$D_{females}$	0.0251**	0.0159	-0.0312**	-0.0129	-0.00554
	(0.011)	(0.015)	(0.016)	(0.013)	(0.008)
N	43,069	43,069	43,069	43,069	43,069
Controls	Y	Y	Y	Y	Y
Income controls	Y	Y	Y	Y	Y

[*] Adds the coefficients of $Twin * OS$ and $Twin * OS * Female$, equivalent to $\beta_6 + \beta_7$.

[1] Estimated using OLS. A description of the set of controls is to be found in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on maternal ID and are in parentheses.

Table A6: Robustness: matching estimators

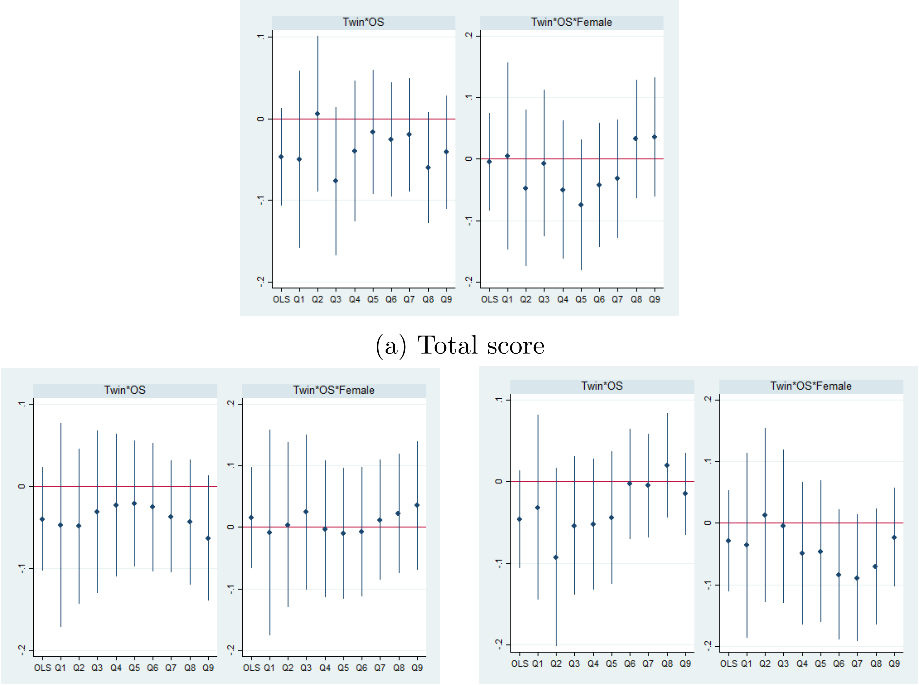
	Total Cito-score			Language score			Math score		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Twin	0.0193	0.0680*	-0.0291	0.0436	0.0738**	-0.00237	-0.00229	0.0665	-0.0558**
	(0.027)	(0.040)	(0.025)	(0.028)	(0.035)	(0.027)	(0.027)	(0.045)	(0.025)
OS	0.0168	0.0731	-0.00397	0.0191	0.0586	0.00471	0.00303	0.0755	-0.0182
	(0.034)	(0.048)	(0.030)	(0.036)	(0.044)	(0.033)	(0.033)	(0.054)	(0.029)
Sex	-0.0491	-0.0187	-0.0842***	0.214***	0.239***	0.187***	-0.354***	-0.314***	-0.402***
	(0.036)	(0.049)	(0.032)	(0.037)	(0.045)	(0.033)	(0.37)	(0.055)	(0.032)
Twin*Female	-0.0605	-0.0986*	-0.0224	-0.0844**	-0.112**	-0.0438	-0.0266	-0.0782	0.0177
	(0.039)	(0.052)	(0.036)	(0.039)	(0.047)	(0.037)	(0.040)	(0.058)	(0.037)
OS*Female	0.00270	-0.0579	-0.00524	-0.00334	-0.0486	-0.0106	0.0194	-0.0505	0.0229
	(0.046)	(0.059)	(0.041)	(0.048)	(0.056)	(0.044)	(0.048)	(0.065)	(0.042)
Twin*OS	-0.0582	-0.119**	-0.0358	-0.0525	-0.0941**	-0.0286	-0.0556	-0.133**	-0.0309
	(0.037)	(0.050)	(0.035)	(0.039)	(0.047)	(0.038)	(0.037)	(0.056)	(0.035)
Twin*OS*Female	-0.0226	0.0459	-0.0281	0.00113	0.0520	-0.00653	-0.0436	0.0342	-0.0635
	(0.050)	(0.062)	(0.048)	(0.052)	(0.060)	(0.050)	(0.052)	(0.069)	(0.049)
D_{males}*	0.0582	0.119**	0.0358	0.0525	0.0941**	0.0286	0.0556	0.133*	0.0309
	(0.037)	(0.050)	(0.035)	(0.039)	(0.060)	(0.038)	(0.037)	(0.056)	(0.035)
$D_{females}$*	-0.0808**	-0.0726*	-0.639*	-0.0514	-0.0421	-0.0351	-0.0992**	-0.0991**	-0.0945**
	(0.037)	(0.041)	(0.036)	(0.038)	(0.040)	(0.036)	(0.040)	(0.044)	(0.038)
N	43,068	43,069	33,030	43,068	43,069	33,030	43,068	43,069	33,030
Kernel M	Y	N	N	Y	N	N	Y	N	N
Inverse Prob.	N	Y	Y	N	Y	Y	N	Y	Y

* Significance refers to outcomes of the Wald-test on the significance of both coefficients β_6 and β_7 combined.

[1] Estimated using OLS. The set of controls is the same as in Table 5.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on the mother's ID and are in parentheses.

[3] Propensity scores based on age, birth order, non-native indicator, household type, whether the mother was in DI in the year of giving birth, household income (age 4), mother working (age 4), mean Cito-score of the school the child is attending.



(a) Total score

(b) Reading score

(c) Math score

Figure A1: Quantile regression, and 95% confidence interval