I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11936

# Biology and the Gender Gap in Educational Performance: The Role of Prenatal Testosterone in Test Scores

Anne C. Gielen
Esmée S. Zwiers

NOVEMBER 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11936

# Biology and the Gender Gap in Educational Performance: The Role of Prenatal Testosterone in Test Scores

**Anne C. Gielen**
*Erasmus University Rotterdam, Tinbergen Institute and IZA*

**Esmée S. Zwiers**
*Erasmus University Rotterdam, Tinbergen Institute*

NOVEMBER 2018

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## Biology and the Gender Gap in Educational Performance: The Role of Prenatal Testosterone in Test Scores*

This paper explores the contribution of biological factors in explaining gender differences in educational performance, with a particular focus on the role of prenatal testosterone. We exploit the fact that prenatal testosterone is hypothesized to transfer in-utero from a male twin to his twin sibling causing exogenous variation in exposure to prenatal testosterone in twins. By using Dutch administrative data and controlling for potential socialization effects, we find that girls with a twin brother score 7% of a standard deviation lower on math compared to girls with a twin sister. Adherence to traditional gender norms can explain this finding, implying that our results are not just driven by biology but materialize depending on environmental factors.

**JEL Classification:** I20, J16

**Keywords:** gender performance gap, twins, prenatals tesosterone

**Corresponding author:**
Anne C. Gielen
Erasmus School of Economics
Erasmus University Rotterdam
PO Box 1738
3000 DR Rotterdam
The Netherlands
E-mail: gielen@ese.eur.nl

* The authors wish to thank the Centraal Instituut voor Toetsontwikkeling (CITO) and Perined for providing access to their data. This paper benefited from comments made by Thomas Buser, Henrik Cronqvist, Gordon Dahl, Guido Imbens, Jan Kabátek, Sacha Kapoor, Olivier Marie, David Neumark, Hannes Schwandt, Zahra Siddique, Andrea Terei, Dinand Webbink, Bas ter Weel, Alice Wu, Basit Zafar, and participants in various conferences and workshops.

# 1 Introduction

Although there has been a quick reversal of the gender gap in educational attainment in the U.S. and most other developed countries in the last decades (e.g. Goldin et al., 2006; Goldin, 2014), this increasing female college attainment stands in sharp contrast with the gender gap in educational test scores, which has remained remarkably stable over time. Generally, boys outperform girls in mathematics (Fryer and Levitt, 2010; Bharadwaj et al., 2015), but fall behind in the reading domain compared to girls (Halpern et al., 2007; Guiso et al., 2008; Banda et al., 2010). These differences are important since test scores typically influence the type of (high) school a child attends, and subsequently influence the type of college one enrols for (Buser et al., 2014; Banda et al., 2010; Ceci et al., 2009), ultimately leading to gender-related earnings differentials.[1] In fact, math skills may become even more important in the labor market due to recent advances in math-intensive technologies (Lippmann and Senik, 2018).[2] Earlier literature has shown that gender differences in math and reading ability can arise from social conditioning and gender-biased environments (e.g. Wilder and Powell, 1989; Miller and Halpern, 2014; Lippmann and Senik, 2018; Reardon et al., 2018). This paper adds biological factors as a potentially additional important driver of gender gaps in educational performance. If there is a role for biological factors in causing such gender differences, ignoring these implies that the role of any discriminatory or gender-biased environmental factors is currently being over-estimated in the literature. Hence, more knowledge on the role of biology is essential, especially in the light of recent policies aiming to promote females in STEM fields of study and STEM careers.

This paper explores biological factors as a potentially additional explanation for gender differences in math and reading performance in childhood. We specifically focus on the role of prenatal testosterone, which is a likely and often mentioned explanation for various gender differences. 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).[3] Little is known to what extent these differences 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. Measuring prenatal testosterone directly in human fetuses is impossible due to practical and ethical constraints. We circumvent this by ex-

[1] For an overview of the literature, trends and explanations of the gender pay gap consult Blau and Kahn (2000), and Blau and Kahn (2017).

[2] Mathematics performance is shown to be related to higher earnings (Altonji, 1995; Arcidiacono, 2004; Joensen and Nielsen, 2009; Altonji et al., 2012; Blau and Kahn, 2017).

[3] 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.

ploiting the twin testosterone transfer (TTT) hypothesis. 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 might transfer in significant concentrations from a male twin to his female uterus mate. This TTT would imply that individuals with a male co-twin are exposed to higher levels of prenatal testosterone than individuals with a female co-twin. Previous studies from other scientific disciplines have used TTT and their findings 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).[4] Since these male-typical cognitive capabilities, e.g. spatial skills, that result from more masculine wiring of the brain are known to be related to boys' advantage in math (Niederle and Vesterlund, 2010), we expect to observe higher math scores for individuals with a male twin than for those with a female twin. In this paper we argue that twinning is a plausible natural experiment to proxy exposure to prenatal testosterone, and that it can be used to identify the effect of elevated prenatal testosterone exposure on math and reading test scores.

Earlier applications of 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. Both of these studies focus on outcomes in adulthood, but the effects of TTT might well appear much earlier in life already. This paper focuses on the role of TTT on outcomes during childhood, in particular educational performance in primary school. We use Dutch administrative data from Statistics Netherlands where we observe all twins born between 1993 and 2003, combined with 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.

To study the causal effect of TTT on test scores we compare children with an opposite-sex twin sibling with children that have a same-sex twin sibling. We control for socialization effects of growing up with a same-sex or opposite-sex sibling by using a control group of closely spaced singletons (CSS) which are siblings whose birth dates are at most twelve months apart.[5] When socialization is similar for twins and CSS, this identification gives the causal effect of TTT on test scores. Our baseline results show that girls with an opposite-sex twin sibling score on average about 7% of a standard deviation lower on math as compared to girls with a twin sister and after controlling for socialization, whereas null effects are found on an aggregate and a reading score. A further investigation in

[4] For males with a male co-twin no evidence for increased masculine behavior or characteristics is found (Resnick et al., 1993; Peper et al., 2009; Tapp et al., 2011; Cronqvist et al., 2015).

[5] The results are robust to using broader windows of 18, 24 and 36 months.

potential mechanisms and explanations for this effect highlights that the effect appears to be concentrated among children growing up in families and areas with more traditional gender norms, and we hypothesize that adherence to the social norm plays an important role here. If TTT causes children to feel different from the typical gender norm, a behavioral response may arise which can offset any potential effect of TTT on test scores. We conclude from this that our findings are not just driven by biological factors, but that the influence of biological factors also strongly depends on environmental factors.

The remainder of this paper proceeds as follows. The next section summarizes the literature on the gender gap in math and reading test scores, and the potential role of prenatal testosterone herein. Section 3 outlines the identification strategy. The data and results are presented in sections 4 and 5. These are followed by a discussion of potential underlying mechanisms in section 6, and a conclusion in section 7.

## 2 Prenatal testosterone and the gender math gap

Several studies for various countries have shown that on average boys perform better in math than girls (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). 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 their performance (Halpern et al., 2007; Machin and Pekkarinen, 2008). The latter implies that boys are more often in both the high and low end of the performance distribution.

Gender differences in educational performance are attributed to both (1) biological differences (i.e. differences in brain development or testosterone exposure) or to (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 threats (e.g. Spencer et al., 1999; Stoet and Geary, 2012; Nollenberger et al., 2014), gender biased environments (Fryer and Levitt, 2010; Bharadwaj et al., 2015), and gender identity norms (Lippmann and Senik, 2018; Reardon et al., 2018). However, our understanding of biological factors explaining gender differences in educational performance is still very limited.

It is well known that early life environments are 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 an important 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. In this period, the fetus is very sensitive to -amongst others- maternal smoking, 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 considers the impact of prenatal exposure to testosterone on educational performance in childhood.

## 2.1 The role of prenatal testosterone

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 for males take place 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 continues until approximately week twenty-four. It is known to be responsible for the development of the testes (Tapp et al., 2011), but this period of gonadal development is also supposed to be critical for the development of the fetal brain (Van de Beek et al., 2004).[6] More specifically, prenatal testosterone is said to 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).[7] To the extent that male-typical cognitive capabilities wired in the brain are responsible for the boys' advantage in math, prenatal testosterone exposure might explain the gender gap in test scores on math and reading.

### 2.1.1 Proxies for prenatal testosterone

The best measure for prenatal testosterone is fetal serum, but direct measurements are infeasible due to the risks it brings to the unborn fetus. Other proxies, like maternal serum testosterone, umbilical cord serum, and amniotic fluid concentrations all have their own disadvantages (Van de Beek et al., 2004). It is for this reason that some direct tests of TTT, involving these proxies, may find conflicting evidence. 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

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

[7] Although the female fetus begins to develop ovaries around week seven of gestation, these ovaries produce only very low levels of estrogens. Estrogens are mainly produced by the maternal placenta, exposure to estrogen levels is similar for both males and female fetuses.

control women (Puts et al., 2008). Disadvantages of using clinical samples are the usually small sample sizes, and limited external validity (Baron-Cohen et al., 2004).

The 2D:4D ratio (the ratio of lengths of the index finger 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), 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 very noisy biomarker as digit ratios would be more correlated with ethnicity than with gender (Cohen-Bendahan et al., 2005).

### 2.1.2 Twin testosterone transfers

Due to the difficulties associated with finding a reliable statistic that measures prenatal exposure to testosterone, more recent studies have started to proxy prenatal testosterone exposure using a sample of twins. Based on evidence with mammals, humans with a male co-twin are hypothesized to be exposed to high levels of prenatal androgens, since testosterone transmits in-utero across amniotic membranes during gestation. This twin testosterone transfer (TTT) hypothesis can be exploited as a natural experiment given that the gender of the co-twin is random (Tapp et al., 2011).

The existence of TTT was first documented in animal-studies, where female rodents with a position near their brothers in the womb were found to 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 TTT among humans is very difficult since direct manipulation of prenatal testosterone levels in human fetuses is clearly unethical (Cohen-Bendahan et al., 2005). Twin studies, however, 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 TTT 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 might 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), however, argue that the effect is less obvious for males, as males themselves are already exposed to relatively high levels of prenatal testosterone.

We use TTT as a proxy for prenatal testosterone exposure. To the best of our knowledge, 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. Both of these papers focus on gender differences in adulthood. However, these difference might originate from gender differences already earlier in childhood. This paper is the first application of TTT to gender differences in educational outcomes during childhood, which likely influence other economic outcomes later in adulthood.

## 3 Empirical strategy

This paper exploits gender variation in twin pairs to examine the causal effect of prenatal testosterone resulting from TTT on test scores. In order to do this, three assumptions must hold: (1) there is a testosterone transfer in humans from a male fetus to the adjacent fetus, (2) the gender distribution is random among and within twin pairs, and (3) there are no confounding factors related to the gender composition of the twin pair that can affect educational outcomes of children in ways other than through a testosterone transfer.[8]

Although direct tests of the first assumption in humans are not available, direct testing on animals showed that in-utero testosterone transfers exist (for an overview see Cohen-Bendahan et al., 2005). This evidence has been used to hypothesize that this testosterone transfer also applies to human twins (Miller, 1994), and has been supported by indirect evidence showing 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). Since no effects are found for males with a male co-twin, possibly as they already have a 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

[8] Our identification strategy follows closely that in Gielen et al. (2016). We refer to their paper for a more detailed discussion on these assumptions.

investigations.

The second identifying assumption is that the gender distribution is random among and within twin pairs. This implies that the gender of a twin sibling is randomly determined. Twins can be monozygotic (identical), when one fertilized egg splits into two same-sex fetuses, or dizygotic (fraternal), when two fertilized eggs develop into two same-sex or opposite-sex fetuses. Identical twins are found to have lower sex ratios than fraternal twins[9], 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. Although this suggests that identical twins are more likely to have a sister (and be female themselves), we are not aware of any evidence that suggests that the probability of being an identical twin is itself determined by levels of prenatal testosterone. For fraternal twins it is commonly assumed that there is an equal probability to be male or female. However, there is evidence showing that fraternal twins are in fact slightly more likely to be male. James (2010) suggests this may be due to higher maternal levels of steroid hormones (testosterone and estrogen) at conception. Maternal serum testosterone levels are found not to be a good proxy for actual prenatal testosterone (Van de Beek et al., 2004; Cohen-Bendahan et al., 2005), but even if maternal and fetal testosterone levels would interact this would only strengthen our identification strategy as individuals with a male co-twin would be exposed to even higher levels of prenatal testosterone (Gielen et al., 2016).

The third assumption stresses that the gender of the co-twin does not influence educational outcomes in any way other than through the prenatal testosterone transfer. This assumption is likely violated as growing up with a brother is different from growing up with a sister, and any such socialization effects resulting from gender-specific parent and/or sibling interactions might also cause the sibling's gender to potentially affect educational outcomes (Peter et al., 2018).[10] To control for this, we define a control group of closely spaced singletons (CSS), consisting of singletons who have a sibling born within 12 months of their own birth date.[11] Provided that any sibling socialization effects are similar for twins and for singletons in the CSS sample[12], any remaining differences in the effects of sibling gender between these two groups can be attributed to the effect of prenatal testosterone exposure.

The control group of CSS 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). Al-

[9] Sex ratios represent the number of boys born for every one hundred girls. Gielen et al. (2016) find a sex ratio of 94.2 for identical twins using data from James (2010).

[10] Similarly research shows that sibling gender can affect women's labor market outcomes (Cools and Patacchini, 2017; Brenøe, 2018).

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

[12] Evidence in favor of this assumption is provided by Björklund and Jäntti (2012), who find strongest sibling correlations for years of schooling among dizygotic twins, those for closely spaced siblings (defined as birth within four years) are stronger and more similar to these dizygotic twins as compared to siblings born more than four years apart.

though the close spacing between siblings in the control group is likely to ensure a socialization closely resembling that between twins, we perform several robustness checks in section 5.1 to assert that there is no evidence for any differential socialization between twins and CSS. Second, the gender of a singleton sibling should not be related to the level of prenatal testosterone. In general, singleton sex ratios can be considered exogenous to prenatal levels of testosterone (see also the discussion in Gielen et al., 2016). However, it is important to note that prenatal testosterone in male singletons is known to decline 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). In this case, second-born singletons in a CSS-pair may experience lower levels of prenatal testosterone in utero. As a robustness check, we estimate the model using only first-borns to assert that this potential concern does not influence our results.

Preferably we would want to distinguish between monozygotic and dizygotic twins (see e.g. Peter et al., 2018), but unfortunately our data does not include information on zygocity. We have to rely (like most other twin studies) on the equal environments assumption (EEA), which states that there are no systematic differences in the environments in which identical and fraternal twins are being raised. The implication of this EEA is that any socialization effects are similar for identical and fraternal twins. Clearly, 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. Yet, earlier studies have shown that the EEA is not violated for spatial ability (Derks et al., 2006) and in several other areas of interest (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), which gives credence to our approach.

The model we estimate to determine the effect of having an opposite-sex twin is displayed in equation 1, and is based on a sample of twins and closely spaced singletons. The variables of interest ($y_{it}$) include an overall test-score, and sub-scores in the domains of math and reading for each individual $i$. We add a female indicator ($female_i$), an indicator for being part of a twin-pair ($twin_i$), an indicator for being part of an opposite-sex sibling pair ($OS_i$), their respective interactions, as well as a vector $\mathbf{X}_{it}$ including other individual and family characteristics, to control for the fact that twins and CSS might have different characteristics and might be born in different types of families, and a series of year dummies. Finally, $u_{it}$ is the individual-specific error term, which is clustered on the maternal identification number.

$$\begin{aligned} y_{it} = \beta_0 + \beta_1 female_i + \beta_2 OS_i + \\ \beta_3 twin_i + \beta_4(female_i * OS_i) + \beta_5(twin_i * female_i) + \\ \beta_6(twin_i * OS_i) + \beta_7(female_i * OS_i * Twin_i) + \mathbf{X}_{it}\delta + u_{it} \end{aligned} \tag{1}$$

In this standard difference-in-difference-in-differences (DDD) model the average difference in test scores between opposite-sex and same-sex twin boys is $D_{twin|male} = \beta_2 + \beta_6$, and the average difference in test scores between opposite-sex and same-sex closely spaced singleton boys is $D_{CSS|male} = \beta_2$. As a result, the double difference for boys is represented by $DD_{male} = \beta_6$. Similarly, for girls the average difference in test scores between opposite-sex and same-sex twins is $D_{twin|female} = \beta_2 + \beta_4 + \beta_6 + \beta_7$, and the average difference in test scores between opposite-sex and same-sex closely spaced singleton girls is $D_{CSS|female} = \beta_2 + \beta_4$. Hence, the double difference for girls equals $DD_{female} = \beta_6 + \beta_7$. The double-difference estimators give the effect of having an opposite-sex twin as compared to having a same-sex twin, after correcting for socialization by subtracting the difference between having a brother and having a sister with the CSS sample. Hence for girls (boys) it gives the effect of having a twin brother (sister) versus having a twin sister (brother), and controls for the effect of having a brother (sister) versus having a sister (brother). If TTT leads to a masculanization of brain structure, we expect to find a positive effect for $DD_{female}$ as girls with an opposite-sex twin sibling would be exposed to higher levels of prenatal testosterone.

# 4 Data

## 4.1 Dutch twins

This paper uses administrative data from Statistics Netherlands covering all registered inhabitants of the Netherlands.[13] We compile our data by matching individuals across the various datasets by their Random Identification Number (RIN), the Dutch (coded) equivalent of the U.S. social security number. We start with the Parent-Child dataset, which matches children to any living parent in the period 1995-2015. From the original information on $15,860,240$ individuals we drop stillbirths ($N = 22,290$) and individuals whose RIN is coded as missing ($N = 547,350$). Siblings are defined as all children born from the same mother.

We merge demographic information from the Municipal Population dataset (in Dutch: Gemeentelijke Basisadministratie, GBA), which contains information on the individuals' year and month of birth, the parents' year and month of birth, gender, and country of origin. We drop individuals who cannot be identified in the Municipal Population dataset ($N = 6,342$) and individuals who are coded as having 15 siblings or more via either parent ($N = 2,090$). First, we select individuals born in the period 1993-2003, as we only observe educational outcomes for these cohorts (more information on educational outcomes is provided in section 4.2). This leaves us with $N = 2,341,814$ observations. Second, we identify twins (or higher order multiples) as siblings with the same birth date, and closely spaced singletons (CSS) as singletons with siblings whose birth dates are within 12 months of an individual's own birth date. The distribution of family structures 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%).[14] We proceed with a sample of twins and CSS, dropping singletons

[13] These data can be accessed through a remote-access facility after a confidentiality agreement has been signed.

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

without siblings, singletons with siblings born outside the 12 month range, and higher order multiples.

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 $\leq$ 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, whose mother can be identified in the data, and who have less than 15 siblings through either parent.

We define a sibling pair as same-sex if the sibling is of the same sex as the individual, and opposite-sex otherwise. In families where there are three (or more) CSS in one family (only $N = 1,760$), it is difficult to classify the sex composition of a sibling pair. We drop these individuals from our sample. Also closely spaced singletons whose birth dates are within 7 months from one another are dropped from the sample ($N = 251$). The distribution of twins and CSS by gender composition is shown in the first columns of Table 2.[15]

## 4.2 Educational outcomes

Data on primary school test-scores is obtained from a high-stakes standardized test performed in the eighth and final grade of elementary education (Cito-test). Note that schools had to give permission to transfer test-scores to Statistics Netherlands, therefore we only observe educational outcomes for those children attending schools who gave permission.[16] The data cover the years 2006 to 2014.[17] For children having multiple test-score records in this period (e.g. due to class retention) the most recent score is preserved. When merging the test-score data to our sample of twins and CSS, we are left with a sample of $50,966$ individuals, as can be seen in the last two columns of Table 2.

The standardized test incorporates performance measures for language, math, information processing, and world orientation.[18] The scores on the various

[15] The twins-sample contains 65.7% same-sex and 34.3% opposite-sex pairs born from 1993 to 2003. Although information on zygosity is unavailable, 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; Fellman and Eriksson, 2006), implying that approximately 68.6% of the twins in our sample are dizygotic.

[16] We observe Cito-scores for approximately 50% of all children born between 1993 and 2003. Missing information can arise from the fact that the child did not take the Cito-test, the child was attending a school that did not take the Cito-test (more than 80% of all schools in the Netherlands administer the Cito-test (Chorny et al., 2010)), or the child did take the Cito but the school did not give permission to transfer the test-scores to Statistics Netherlands.

[17] Test scores for 2015 are available but are not being used as the structure of the test changed in 2015 and hence scores are not comparable to those in earlier years.

[18] The questions on world orientation are optional and hence not completed by all children.

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: Sample of twins and closely spaced singletons (CSS). The first column shows the distribution of opposite-sex (OS) and same-sex (SS) pairs in the overall GBA. The second panel shows the same distributions for the sample of individuals for whom we observe test scores in the data.

(sub)parts are translated into an aggregated score ranging between 501 and 550. In order to be able to compare scores across different years (and hence different tests), the aggregate score and the sub-scores for math and reading are standardized by year in a Z-score.[19]

### 4.3 Descriptive statistics

Average standardized test scores differ between boys and girls, and between twins and CSS (Table 3).[20] Boys outperform girls in math, and girls perform significantly better in reading. This gender-specific pattern in performance gaps is consistent with the general pattern found in the literature (see e.g. Guiso et al., 2008; Fryer and Levitt, 2010; OECD, 2015), and it is visible for both the full sample and for the sub-samples of twins and closely spaced singletons. For twins the gender gaps in school performance are even more pronounced.

Table 4 shows that gender gaps in test performance also vary with the gender of one's sibling.[21] Although we observe no significant differences in test scores between opposite-sex and same-sex closely spaced singletons, girls in opposite-sex twin pairs score significantly lower in math and the aggregate score as opposed to same-sex twin girls. If anything, this is suggestive evidence against the

[19] Z-scores for individual $i$ in year $t$ are defined as Z-score$_{it}$=(score$_{it}$-$\mu_t$)/$\sigma_t$, where score$_{it}$ denotes the test (sub-)score, $\mu_t$ denotes the average test (sub-)score in year $t$, and $\sigma_t$ denotes the standard deviation in (sub-)scores in year $t$.

[20] Exact variable definitions are provided in Appendix Table A1.

[21] Gender differences in the distribution of test scores are presented in Figure A1 and Figure A2.

Table 3: Gender gaps in test performance

| | All children | | | Sample of twins and CSS | | |
|---|---|---|---|---|---|---|
| | 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: Test scores are standardized with mean zero and standard deviation one.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TTT hypothesis. Note however that this simple comparison neglects potential socialization effects, as well as the impact of other background characteristics. For example, opposite-sex and same-sex twins differ significantly in their family background, where opposite-sex twins are born from slightly older parents and are raised in somewhat smaller families. These differences could hint at a preference that parents may have for children of mixed genders (e.g. Angrist and Evans, 1998).

There are also marked differences between twins and CSS. Twins have slightly higher test scores than CSS[22], which is at least partly due to their different family background. Higher educated mothers are more likely to built a career before having children. Since twinning probabilities increase with maternal age (Rosenzweig and Wolpin, 1980; Bronars and Grogger, 1994; Jacobsen et al., 1999) and the use of artificial reproductive technologies (ART) (Bhalotra et al., 2016), we observe that twins are born to older mothers (and fathers) in high income families.[23] This also explains why twins have a lower parity on average. Our empirical approach in the next section accounts for these differences in family background when estimating the effect of the gender of a twin sibling on educational test performance.

## 5 Results

The results from the baseline specification for the aggregate test score are presented in Table 5. The twin coefficient is positive and significant in the specification without controls (column 1), and becomes smaller and insignificant once

[22] Related to this, twins have a lower age at test, as the flip-side of better school performance is a lower probability of repeating a grade.

[23] Household income - i.e. the sum of the earnings of both parents in a particular year - is measured in the year the child turns 4 years old due to income information only being available from 1999 onwards. In the Netherlands, children start elementary school at age 4. We do not observe income information for children born before 1995, which explains the lower number of observations for this variable.

Table 4: 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* | | | | | | | | |
| Total score (Std) | -0.088 | -0.055 | -0.238 | -0.253 | -0.009 | *** | ** | |
| Language (Std) | 0.057 | 0.066 | -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 | *** | | |
| Parity (birth order) | 1.735 | 1.743 | 2.106 | 2.130 | 1.806 | *** | | |
| Spacing | 0 | 0 | 11.483 | 11.490 | | *** | | |
| Non-native (dummy) | 0.158 | 0.166 | 0.382 | 0.421 | 0.211 | *** | | *** |
| Family size | 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 (dummy) | 0.020 | 0.016 | 0.019 | 0.015 | 0.013 | | ** | |
| *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=7,608* | *N=12,601* | *N=2,995* | *N=2,839* | *N=641,882* | | | |
| | | | | | | | | |
| HH-income (at age 4)* | 44,023.21 | 43,014.93 | 32,906.84 | 31,706.77 | 41,144.33 | *** | * | |
| Mother works (dummy)* | 0.658 | 0.664 | 0.498 | 0.499 | 0.671 | *** | | |
| | *N=6,552* | *N=10,660* | *N=2,513* | *N=2,314* | *N=543,672* | | | |

| | 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* | | | | | | | | |
| Total score (Std.) | 0.042 | 0.037 | -0.188 | -0.189 | 0.039 | *** | | |
| Language (Std.) | -0.071 | -0.077 | -0.356 | -0.351 | -0.079 | *** | | |
| Math (Std.) | 0.174 | 0.185 | 0.071 | 0.074 | 0.185 | *** | | |
| Age at test (Months) | 12.067 | 12.108 | 12.125 | 12.114 | 12.037 | *** | *** | |
| Parity (birth order) | 1.730 | 1.756 | 2.138 | 2.137 | 1.805 | *** | * | |
| Spacing | 0 | 0 | 11.481 | 11.490 | | *** | | |
| Non-native (dummy) | 0.158 | 0.173 | 0.397 | 0.372 | 0.210 | *** | *** | * |
| Family size | 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 (dummy) | 0.020 | 0.016 | 0.017 | 0.011 | 0.012 | | * | * |
| *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=7,193* | *N=12,039* | *N=2,805* | *N=2,886* | *N=636,303* | | | |
| | | | | | | | | |
| HH-income (at age 4)* | 44,973.46 | 43,344.22 | 32,484.99 | 33,062.50 | 41,610.28 | *** | | |
| Mother works (dummy)* | 0.668 | 0.679 | 0.498 | 0.520 | 0.665 | *** | | |
| | *N=6,147* | *N=10,151* | *N=2,315* | *N=2,417* | *N=535,643* | | | |

[*] 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

controls are added (column 2). This clearly shows that twins and CSS are born into different families. These results remain unchanged once we focus on the smaller sample for which family income information is available (columns 3-5). The dummy variable for having an opposite-sex sibling is not significant in any of the specifications, suggesting limited to no role for socialization effects as sibling gender by itself does not affect educational outcomes. The female indicator consistently shows that girls have significantly lower aggregate test scores than boys (by approximately 5% of a standard deviation and conditional on characteristics $\mathbf{X}_i$).

The effects of opposite-sex twinning for boys ($DD_{male}$) and for girls ($DD_{female}$) are not significantly different from zero. If anything, the effect for girls is negative suggesting that females with a male uterus-mate would perform about 5% of a standard deviation worse on the aggregate score, when controlling for the socialization effect of growing up with a brother. This effect is contrary to what would be expected from the TTT hypothesis, but might mask differential effects for math and reading.

The results for the reading and math sub-scores are shown in the left and right panel of Table 6, respectively. Twins appears to have higher math and reading scores than CSS, but these differences disappear once we include relevant controls for family background. The opposite-sex sibling dummy is insignificant in all specifications. The gender dummy reveals that girls have a significant advantage in reading (2% of a standard deviation), whereas boys have an advantage in the math-domain (about 4% of a standard deviation). We find no significant effect for opposite-sex twinning on reading scores for either boys and girls. However, for math scores we find that girls with a twin brother perform about 7% of a standard deviation worse, even after controlling for socialization effects and family background.[24]

The results in Tables 5 and 6 focus on mean test scores, but previous research has shown evidence for the presence of gender differences in test-score distributions (Halpern et al., 2007; Machin and Pekkarinen, 2008). To check for any such effects, we also estimate quantile regression models, but these results are very similar to the OLS estimates as can be seen in Figure A3.

The negative effect for girls with an opposite-sex twin ($DD_{female}$) on math might seem counter-intuitive as the TTT hypothesis would predict that girls with a twin brother are exposed to higher concentrations of prenatal testosterone, and hence would display improved math performance (and potentially worse reading performance). We do not find evidence for this, nor do we find any effect of opposite-sex twinning for boys. In section 6 we discuss various explanations for our findings.

[24] Table A2 and Table A3 show that the results are robust to estimating the models separately for boys and girls.

Table 5: Results for aggregate test score (standardized)

| | Aggregate score | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Twin | 0.226*** | -0.006 | 0.220*** | -0.002 | -0.007 |
| | (0.025) | (0.021) | (0.026) | (0.023) | (0.023) |
| OS | 0.001 | 0.010 | -0.006 | 0.004 | 0.005 |
| | (0.030) | (0.025) | (0.032) | (0.027) | (0.027) |
| Female | -0.064** | -0.068*** | -0.065* | -0.069** | -0.067** |
| | (0.032) | (0.026) | (0.034) | (0.028) | (0.028) |
| Twin*Female | -0.029 | -0.041 | -0.039 | -0.040 | -0.039 |
| | (0.035) | (0.029) | (0.038) | (0.031) | (0.031) |
| OS*Female | 0.014 | -0.027 | 0.015 | -0.017 | -0.018 |
| | (0.039) | (0.033) | (0.042) | (0.036) | (0.036) |
| Twin*OS | 0.004 | -0.040 | -0.020 | -0.047 | -0.047 |
| | (0.034) | (0.028) | (0.036) | (0.031) | (0.030) |
| Twin*OS*Female | -0.052 | -0.004 | -0.034 | -0.005 | -0.005 |
| | (0.044) | (0.037) | (0.047) | (0.040) | (0.040) |
| $DD_{male}$ | 0.004 | -0.040 | -0.020 | -0.047 | -0.047 |
| | (0.034) | (0.028) | (0.036) | (0.031) | (0.030) |
| $DD_{female}$ | -0.048 | -0.045 | -0.053 | -0.052* | -0.051* |
| | (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 |

$^{\text{Note:}}$ Results are based on OLS model. 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. Specifications 3-5 report results for a smaller sample, for which information on household income and maternal employment when the child is 4 years old is available. Standard errors are clustered on maternal ID and are in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

## 5.1 Robustness

Although our results in the previous section suggest that girls' math performance is affected by having a twin brother (as opposed to a twin sister), we should treat these results with care. There are several potential concerns with our identification strategy, that might lead to misinterpretations. In this section, we will discuss each of these and examine the impact they may have on our results.

One potential concern for our identification is that maternal levels of testosterone are known to be lower if spacing between subsequent children is less than four years (Maccoby et al., 1979; Baron-Cohen et al., 2004). To address this issue we restrict the sample to first born children only. This approach also deals with some potential concerns about the validity of CSS as an appropriate control group. First, taking first borns takes into account that the decision to have a second child may be endogenous to the gender of the first child (Dahl and

Table 6: Results for standardized reading and math score

| | Reading score | | | | | Math score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (1) | (2) | (3) | (4) | (5) |
| Twin | 0.273*** | 0.026 | 0.267*** | 0.027 | 0.022 | 0.112*** | -0.042** | 0.114*** | -0.033 | -0.036 |
| | (0.026) | (0.022) | (0.028) | (0.024) | (0.023) | (0.023) | (0.021) | (0.025) | (0.023) | (0.023) |
| OS | -0.005 | 0.003 | -0.004 | 0.005 | 0.006 | -0.003 | 0.006 | -0.016 | -0.006 | -0.006 |
| | (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 | -0.061* | -0.074** | -0.074* | -0.078** | -0.077** | 0.014 | 0.009 | 0.009 | 0.014 | 0.014 |
| | (0.036) | (0.030) | (0.039) | (0.032) | (0.032) | (0.034) | (0.030) | (0.037) | (0.032) | (0.032) |
| OS*Female | 0.024 | -0.016 | 0.015 | -0.019 | -0.020 | 0.008 | -0.020 | 0.024 | 0.003 | 0.002 |
| | (0.040) | (0.034) | (0.043) | (0.037) | (0.037) | (0.038) | (0.034) | (0.041) | (0.037) | (0.037) |
| Twin*OS | 0.012 | -0.030 | -0.014 | -0.040 | -0.040 | -0.008 | -0.042 | -0.026 | -0.046 | -0.046 |
| | (0.035) | (0.029) | (0.038) | (0.032) | (0.032) | (0.032) | (0.028) | (0.034) | (0.031) | (0.031) |
| Twin*OS*Female | -0.040 | 0.007 | -0.016 | 0.015 | 0.015 | -0.055 | -0.019 | -0.049 | -0.029 | -0.029 |
| | (0.044) | (0.038) | (0.048) | (0.042) | (0.042) | (0.043) | (0.039) | (0.046) | (0.042) | (0.042) |
| $DD_{male}$ | 0.012 | -0.030 | -0.014 | -0.040 | -0.040 | -0.008 | -0.042 | -0.026 | -0.046 | -0.046 |
| | (0.035) | (0.029) | (0.038) | (0.032) | (0.032) | (0.032) | (0.028) | (0.034) | (0.031) | (0.031) |
| $DD_{female}$ | -0.028 | -0.023 | -0.030 | -0.025 | -0.025 | -0.063* | -0.062** | -0.075** | -0.076** | -0.075** |
| | (0.033) | (0.028) | (0.036) | (0.031) | (0.031) | (0.033) | (0.030) | (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 |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Specifications 3-5 report results for a smaller sample, for which information on household income and maternal employment when the child is 4 years old is available. Standard errors are clustered on maternal ID and are in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Moretti, 2008; Blau et al., 2017), in which case CSS would not make up an appropriate control group. Second, spacing matters for parental time investments. First born children engage significantly more in quality-time activities with their parents (e.g. reading and playing) than later-born siblings (Price, 2008), which can explain the significant effect of birth order on child outcomes found in the literature. Taking a sample of first-borns accounts for these birth order effects, by improving comparability between the group of twins (treated) and CSS (control). The results in Panel B of Table 7 illustrate that the point estimates in this specification are comparable to those in the baseline specification, but the significance for $DD_{female}$ in math scores drops, which is mostly due to a decrease in precision as the number of observations halved.[25,26]

Another potential threat to our identification could be that CSS appear to be an inappropriate control group to capture socialization effects. Our estimator might be biased if socialization effects in families with CSS differ from those in families with twins (according to the gendermix of the sibling pair). We address this potential concern in various ways. First, it is important to stress that the gender of the sibling in a CSS pair does not seem to affect test scores; the results are mainly driven by the differential effect of sibling gender within twin pairs.[27]

Table 4 has shown that households with twins and CSS are different in various characteristics. In particular, the native origin of the family appears to be an important difference, which might affect socialization effects between siblings, e.g. due to differential cultural and religious factors. Furthermore, there might be misreporting in the birth dates of foreign born children which might contaminate the sample of twins or CSS.[28] To check the appropriateness of using CSS as a control group, we limit the sample to children of native Dutch parents. The results in Panel C of Table 7 show that the double difference estimate for girls is larger and significant, whereas the double difference estimate for males is lower and less precise compared to the baseline.[29] Hence, these estimates confirm our main results and, if anything, may suggest that our baseline estimate is somewhat conservative.

Another difference between families with CSS and families with twins is the number of children in a household (see Table 4). Twins are -on average- born in smaller families than CSS, and it might be that socialization between siblings (of different genders) varies between larger and smaller families. In order to further check the appropriateness of using CSS as a control group, we limit the sample to children of two-child families only such that twins and CSS in the sample grow up in families of equal size (Panel D of Table 7). The double difference estimates

[25] The full estimation results are available in Appendix Tables A4 and A5.

[26] It does not matter whether the first born is a boy or a girl, as we find a similar pattern when estimating the model for second borns (results available on request).

[27] This is consistent with Peter et al. (2018) who find no effect of sibling gender on years of schooling for regular siblings and close siblings (defined as birth dates within 24 months). They do find an effect of sibling gender on years of schooling for dizygotic twins (i.e. girls with a twin brother have 0.112 more years of schooling.).

[28] As an example, due to misreporting 20% of the Turkish population has a registered birth date in January (Torun and Tumen, 2016).

[29] The full estimation results are available in Appendix Tables A4 and A5.

Table 7: Robustness results

| | Aggregate score | | Reading score | | Math score | |
|---|---|---|---|---|---|---|
| | $DD_{male}$ | $DD_{female}$ | $DD_{male}$ | $DD_{female}$ | $DD_{male}$ | $DD_{female}$ |
| A. Baseline | -0.047 | -0.051* | -0.040 | -0.025 | -0.046 | -0.075** |
| | (0.030) | (0.030) | (0.032) | (0.031) | (0.031) | (0.032) |
| $N$ | 43,069 | 43,069 | 43,069 | 43,069 | 43,069 | 43,069 |
| B. First born only | -0.033 | -0.057 | -0.022 | -0.036 | -0.051 | -0.067 |
| | (0.050) | (0.048) | (0.052) | (0.048) | (0.050) | (0.052) |
| $N$ | 19,576 | 19,576 | 19,576 | 19,576 | 19,576 | 19,576 |
| C. Natives only | -0.016 | -0.090** | -0.014 | -0.064* | -0.015 | -0.112*** |
| | (0.036) | (0.037) | (0.038) | (0.037) | (0.036) | (0.039) |
| $N$ | 34,003 | 34,003 | 34,003 | 34,003 | 34,003 | 34,003 |
| D. Two-child family only | -0.059 | -0.123** | -0.035 | -0.073 | -0.070 | -0.182*** |
| | (0.055) | (0.055) | (0.057) | (0.055) | (0.056) | (0.059) |
| $N$ | 14,034 | 14,034 | 14,034 | 14,034 | 14,034 | 14,034 |
| E. *CSS window:* | | | | | | |
| 18 months | -0.063*** | -0.047*** | -0.069*** | -0.024 | -0.046*** | -0.061*** |
| | (0.017) | (0.016) | (0.017) | (0.016) | (0.017) | (0.018) |
| $N$ | 132,650 | 132,650 | 132,650 | 132,650 | 132,650 | 132,650 |
| 24 months | -0.065*** | -0.050*** | -0.070*** | -0.030** | -0.052*** | -0.062*** |
| | (0.015) | (0.015) | (0.016) | (0.015) | (0.016) | (0.016) |
| $N$ | 279,980 | 279,980 | 279,980 | 279,980 | 279,980 | 279,980 |
| 36 months | -0.064*** | -0.054*** | -0.069*** | -0.029* | -0.050*** | -0.070*** |
| | (0.015) | (0.015) | (0.016) | (0.015) | (0.015) | (0.016) |
| $N$ | 492,264 | 492,264 | 492,264 | 492,264 | 492,264 | 492,264 |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5 (Column 5). Standard errors are clustered on maternal ID and are in parentheses. Full estimation results can be found in Appendix Tables A4 and A5.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

for boys remains insignificant, whereas those for girls become considerably larger. Overall, these results continue to support our conclusion that having an opposite-sex twin is associated with lower math scores for girls, and they show that our baseline estimate might be rather conservative.

A crucial assumption for the definition of CSS as an appropriate control group is the 12-month window within which CSS are defined. This window is explicitly very narrow as to increase the probability that socialization effects between closely spaced singletons are similar to those between twins, but this comes at a cost of a relatively low number of observations which might decrease the precision of the estimates. Panel E of Table 7 presents the results of a series of estimations in which we investigate how robust our findings are to extended windows within which CSS are defined (i.e. 18 months, 24 months and 36 months, respectively). The double difference estimates for girls are highly robust to using different bandwidths, ranging from 2.4 to 3.0 percent of a standard deviation for reading and ranging from 6.1 to 7.0 percent of a standard deviation for math across the specifications. For boys the estimates are less robust, as they increase in size and significance. The positive double difference effects in math for boys seem to result from increased precision in the estimation. However, for the reading specification the increased significance is likely due to the fact that the gender-specific socialization effects between CSS become different from that between twins when sibling spacing increases, which is reflected by the increasing estimate for having an opposite-sex sibling and the decreasing estimate for having an opposite-sex twin (see Appendix Tables A6 and A7). This suggests that defining CSS using a wider window for birth spacing reduces the suitability of CSS as a control group.

To provide further credibility for the use of CSS as a control group, we employ a matching estimator to make the sample of CSS and twins more comparable. The results using Kernel matching as well as Inverse Probability Matching are presented in Table A8.[30] Although the effects from the matching estimation are larger than the baseline estimates, suggesting the latter are a conservative estimate of the true effect, our overall conclusions remain unchanged.

All in all, we interpret the above results as supportive evidence for the suitability of CSS as a control group. Although we do not find evidence that the gender of a sibling affects test scores of very closely spaced siblings, we cannot completely rule out that socialization is different between CSS and between twins. If our effects for twins would be driven by socialization effects[31], this would imply that parents or teachers would have to differentially invest in the education/training of twins based on the gender of a twin sibling, but would not respond to the sibling gender for singletons.

## 6 Mechanisms at work

The result that girls with a twin brother perform 7% of a standard deviation lower on math seems somewhat counterintuitive, as from the TTT hypothesis one would expect that these girls would be more male-typical and hence their educational performance would also appear as being more male-typical. In this section we investigate four potential mechanisms that may explain our findings.

First, we explore the role that TTT may have on other birth outcomes, that may in turn affect educational outcomes later in life. The medical literature has shown that boys typically have a higher birth weight than girls (Bouckaert et al., 1992; Voldner et al., 2009), and the economic literature has provided evidence that birth weight is a robust predictor of cognitive development and academic outcomes (Autor et al., 2017; Bharadwaj et al., 2018). If sharing the intra-uterine environment with an opposite-sex fetus would affect birth weight through TTT, then this could have a direct impact on educational outcomes later

[30] 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, 3 and 5). Inverse Probability Matching (IPM) is also used, 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 a propensity score between 0.1 and 0.9 (column 2, 4, and 6).

[31] Socialization effects are stressed as important in a related study by Peter et al. (2018).

Table 8: Birth outcomes

| | Birth weight (grams) | Birth weight (grams) | Gestation (days) | Gestation (days) |
|---|---|---|---|---|
| $DD_{males}$ | 66.274*** | 64.586*** | 2.375*** | 2.324*** |
| | (18.235) | (17.903) | (0.537) | (0.535) |
| $DD_{females}$ | 71.525*** | 70.108*** | 2.541*** | 2.428*** |
| | (17.814) | (17.607) | (0.527) | (0.527) |
| N | 80, 663 | 80, 663 | 80, 663 | 80, 663 |
| Controls | N | Y | N | Y |

Notes: Results are based on OLS model. Controls are birth order dummies, maternal age at birth, non-native dummy, and year of birth dummies. Standard errors are clustered on maternal ID and are in parentheses. Full estimation results are available in Appendix Table A9

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in life. Miller and Martin (1995) show that birth weight in mice is higher for females located between two male fetuses as opposed to females located between two female fetuses. For humans, however, the evidence is mixed and inconclusive (Orlebeke et al., 1993; Glinianaia et al., 1998; Loos et al., 2001; Blickstein and Kalish, 2003). Table 8 shows the results from a model in which we look at birth weight as the relevant outcome.[32] We find that girls indeed have lower birth weight than boys, and that birth weight is higher for girls with a twin brother. Evidently, this cannot explain our baseline effects for girls with a twin brother. If anything, the positive birth weight effect for girls with an opposite-sex twin sibling would translate into higher math scores, not lower scores. Furthermore, when looking at gestational age - another important birth outcome - there is more evidence for a positive effect from opposite-sex twinning for girls. Given that the effect on birth weight for males with a twin sister also does not seem to translate into higher test scores later in childhood, we interpret these results as evidence that other birth related outcomes cannot explain our baseline findings.

A second, alternative, explanation could be that TTT wires the brain differently (Jordan-Young, 2010) leading to gender differences in various psychological traits, but that these non-cognitive skills impact educational outcomes in a more complex way than our baseline model allows for. For example, externalizing behavior (e.g. getting angry, fighting, acting impulsively) that is more prevalent among boys is a robust predictor of eight grade suspension (Bertrand and Pan, 2013). If having an opposite-sex twin impacts grade retention, then this might offset any potential impact on test scores. Table 9 shows, however, that opposite-sex twinning does not seem to be related with any measure of grade retention.[33] Furthermore, a potential differential impact of non-cognitive skills

[32] Data on birth outcomes is available from 2004 to 2014 (PRNL dataset). The identification of twins and CSS is exactly the same as described in section 4.1, but we merge the remaining twins and CSS to the data on birth outcomes. Note that these twins and CSS are not the same as observed in the test-score data, as we only have information for individuals born from 2004 to 2014. The procedure leaves a sample of 63, 253 twins and 17, 410 CSS.

[33] We use two proxies for grade retention as a direct measure is unavailable. We use an indicator for having multiple Cito records in the data for retention in the final grade of elementary

Table 9: Other educational outcomes of interest

| | Grade retention | | | Teacher assessment | |
|---|---|---|---|---|---|
| | $DD_{male}$ | $DD_{female}$ | | $DD_{male}$ | $DD_{female}$ |
| | | | School advice: | | |
| I(Multiple records) | -0.001 | 0.001 | - At least lower/general pre-vocational track | 0.011 | -0.012 |
| | (0.003) | (0.002) | | (0.011) | (0.011) |
| $N$ | 43,069 | 43,069 | $N$ | 30,944 | 30,944 |
| I(age $\geq$ 13) | 0.004 | 0.006 | - At least general pre-vocational track | 0.007 | -0.004 |
| | (0.005) | (0.005) | | (0.013) | (0.013) |
| $N$ | 43,069 | 43,069 | $N$ | 30,944 | 30,944 |
| | | | - At least general/higher pre-vocational track | 0.009 | 0.002 |
| | | | | (0.016) | (0.017) |
| | | | $N$ | 30,944 | 30,944 |
| | | | - At least higher pre-vocational track | 0.013 | -0.003 |
| | | | | (0.017) | (0.017) |
| | | | $N$ | 30,944 | 30,944 |
| | | | - At least higher pre-vocational/general track | -0.018 | -0.006 |
| | | | | (0.019) | (0.019) |
| | | | $N$ | 30,944 | 30,944 |
| | | | - At least general track | -0.028 | -0.025 |
| | | | | (0.018) | (0.018) |
| | | | $N$ | 30,944 | 30,944 |
| | | | - At least general/academic track | -0.029 | -0.022 |
| | | | | (0.018) | (0.018) |
| | | | $N$ | 30,944 | 30,944 |
| | | | - At least academic track | -0.042*** | -0.006 |
| | | | | (0.016) | (0.016) |
| | | | $N$ | 30,944 | 30,944 |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. We proxy grade eighth retention with an indicator for having multiple Cito-records in our data. In addition, we proxy any grade retention with an indicator that the child is 13 years or older at the time of taking the test. Standard errors are clustered on maternal ID and are in parentheses. The teacher assessment outcomes are indicators for having a school advice greater or equal to category X. There are nine categories and they range from advice for the lower vocational track (1) to the pre-university track (9).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

on overall school performance is also not reflected in the teachers' assessment of the student's overall ability.[34] Hence, we conclude from this that our negative findings for opposite-sex twinning for girls are not driven by non-cognitive skills that impact other educational outcomes in a way that would offset the impact on test scores.

school. We proxy any grade retention with an indicator variable for being 13 years or older at the time of taking the test.

[34] The teacher assessments of the child's ability is communicated to students by means of a 'school advice' for a secondary school track.

A third, but related, argument is that the impact of non-cognitive skills on educational outcomes may differ by gender. For example, Bertrand and Pan (2013) show that there are gender differences in the non-cognitive returns to parental inputs, and that the non-cognitive development of boys is much more responsive to adverse parental investments resulting from parental divorce than that of girls. Also Autor et al. (2017) and Brenøe and Lundberg (2017) find that family disadvantage disproportionally negatively affects the behavior and school outcomes of boys relative to girls. In Table 10 we investigate how our results vary with the household situation. Strikingly, our findings seem to be concentrated among non-divorced and two-parent households. Hence, differences in how boys and girls deal with adverse shocks in household composition or stability do not seem to explain our negative effects of opposite-sex twinning for girls. In fact, a negative effect of opposite-sex twinning also appears for boys in these "traditional" families. This might be suggestive evidence for the fact that boys with a twin brother receive a 'double dose' of prenatal testosterone (Resnick et al., 1993; Peper et al., 2009), and hence perform better than boys with a twin sister.

A further inspection of various subsamples confirms that the effect of opposite-sex twinning may not be uniformly distributed, but may rather depend on the environment in which a child is being raised.[35] Table 10 shows that children from high income households and children from more advantaged backgrounds are more likely to experience a negative impact of opposite-sex twinning on test scores. This evidence clearly suggests that our results are not purely biological but that they are also strongly subject to environmental influences. Recently research has shown that gender achievement gaps are more pronounced in areas characterized by a higher socioeconomic background of its inhabitants (Reardon et al., 2018). One explanation for this finding is that traditional gender norms may be more stereotypical in these areas, e.g. if the man is the main bread-winner in the family. Lippmann and Senik (2018) find a smaller gender math gap in East Germany and former Soviet countries, areas with more equal gender norms due to socialism, and thereby argue that gender norms are an important determinant of the gender math gap.

An explanation could be that TTT leads to more male-typical characteristics for girls, in behavior as well in morphological attributes (e.g. Cohen-Bendahan et al., 2004; Peper et al., 2009; Vuoksimaa et al., 2010a). And that being (perceived as) different from a typical girl (or boy) - i.e. different from the norm - may have a much larger impact on children if they were raised in a family with traditional gender norms. Hereby, feeling different may give rise to a behavioral response that offsets the potential effect of TTT on math scores. The results in Table 10 provide support for this hypothesis as the effects are concentrated among children growing up in "traditional" families in terms of a two married parents household. Arguably these are also households in which gender norms are strongest (Reardon et al., 2018). However, to get a better understanding of this mechanism we exploit regional differences in traditional gender norms across

[35] The importance of environmental factors has earlier been stressed by e.g. Björklund et al. (2006).

Table 10: By household characteristics

| | Aggregate score | | Reading score | | Math score | |
|---|---|---|---|---|---|---|
| | $DD_{male}$ | $DD_{female}$ | $DD_{male}$ | $DD_{female}$ | $DD_{male}$ | $DD_{female}$ |
| Baseline | -0.047 | -0.051* | -0.040 | -0.025 | -0.046 | -0.075** |
| | (0.030) | (0.030) | (0.032) | (0.031) | (0.031) | (0.032) |
| $N$ | $43,069$ | $43,069$ | $43,069$ | $43,069$ | $43,069$ | $43,069$ |
| | | | | | | |
| *By household type:* | | | | | | |
| Two-parent HH | -0.070** | -0.066** | -0.060* | -0.059* | -0.063* | -0.069* |
| | (0.033) | (0.033) | (0.035) | (0.034) | (0.033) | (0.035) |
| $N$ | $36,404$ | $36,404$ | $36,404$ | $36,404$ | $36,404$ | $36,404$ |
| One-parent HH | 0.068 | -0.011 | 0.076 | 0.115 | 0.020 | -0.134* |
| | (0.076) | (0.075) | (0.077) | (0.075) | (0.078) | (0.080) |
| $N$ | $6,383$ | $6,383$ | $6,383$ | $6,383$ | $6,383$ | $6,383$ |
| | | | | | | |
| *Household stability:* | | | | | | |
| Non-divorced | -0.083** | -0.094*** | -0.064* | -0.078** | -0.085** | -0.095** |
| | (0.035) | (0.035) | (0.037) | (0.035) | (0.035) | (0.037) |
| $N$ | $33,374$ | $33,374$ | $33,374$ | $33,374$ | $33,374$ | $33,374$ |
| Divorced | 0.169** | 0.089 | 0.131 | 0.194** | 0.150* | -0.058 |
| | (0.084) | (0.083) | (0.090) | (0.083) | (0.083) | (0.088) |
| $N$ | $5,370$ | $5,370$ | $5,370$ | $5,370$ | $5,370$ | $5,370$ |
| Not married | -0.062 | 0.071 | -0.078 | 0.055 | -0.032 | 0.060 |
| | (0.095) | (0.091) | (0.097) | (0.092) | (0.098) | (0.097) |
| $N$ | $4,325$ | $4,325$ | $4,325$ | $4,325$ | $4,325$ | $4,325$ |
| | | | | | | |
| *By HH income:* | | | | | | |
| Low-income | -0.010 | -0.030 | -0.012 | 0.009 | -0.005 | -0.074* |
| | (0.038) | (0.037) | (0.040) | (0.037) | (0.039) | (0.039) |
| $N$ | $25,429$ | $25,429$ | $25,429$ | $25,429$ | $25,429$ | $25,429$ |
| High-income | -0.095* | -0.133** | -0.058 | -0.125** | -0.128*** | -0.116** |
| | (0.050) | (0.052) | (0.052) | (0.053) | (0.049) | (0.056) |
| $N$ | $17,640$ | $17,640$ | $17,640$ | $17,640$ | $17,640$ | $17,640$ |
| | | | | | | |
| *By subsidy factor:* | | | | | | |
| Disadvantaged | -0.010 | -0.013 | -0.077 | 0.040 | 0.074 | -0.046 |
| | (0.074) | (0.069) | (0.076) | (0.070) | (0.075) | (0.072) |
| $N$ | $6,352$ | $6,352$ | $6,352$ | $6,352$ | $6,352$ | $6,352$ |
| Non-disadvantaged | -0.028 | -0.080** | -0.006 | -0.059 | -0.052 | -0.102** |
| | (0.036) | (0.037) | (0.038) | (0.037) | (0.036) | (0.039) |
| $N$ | $30,647$ | $30,647$ | $30,647$ | $30,647$ | $30,647$ | $30,647$ |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. A household is high-income if household income at age 4 is greater or equal to the average household income at age 4 of all children observed in the test-score data. Standard errors are clustered on maternal ID and are in parentheses. Full estimation results are available in Appendix Table A10, A11, and A12.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the Netherlands. Specifically, we make a distinction between children living in municipalities in the 'Bible Belt', an area with a high number of conservative Christians, and those living outside this area. Municipalities are defined as more or less religious based on the share of votes for the orthodox Calvinist political party (in Dutch: Staatkundig Gereformeerde Partij, SGP) in the 2017 national parliamentary elections.[36] The SGP is known for its valuation of traditional gender norms, i.e. considering the man as the head of a household. First, we find suggestive evidence for a larger gender math gap in test scores in more religious versus less religious areas (-40.5% of a standard deviation versus -38.6% of a standard deviation).[37] This is consistent with Lippmann and Senik (2018) and Reardon et al. (2018) who find larger gender gaps in areas with more traditional gender norms, and this suggests that more traditional gender norms prevail in more religious areas. Second, Table 11 shows that the double-difference estimators are more negative for girls living in more religious areas.[38] This illustrates that our baseline effects are concentrated among girls living in municipalities that are characterized by more traditional gender norms, suggesting that the effect of biological factors on gender differences in test scores materializes more in more traditional environments. Hence, adherence to a social norm plays an important role here. This argument aligns closely with an earlier study by Gielen et al. (2016), who find a marginal negative earnings effect for females with a twin brother, which the authors explain by labor market discrimination against females with attributes that are perceived as more masculine, i.e. females that do not adhere to the norm. In this study we argue that the feeling of being different, due to TTT, might give rise to a behavioral response, which may impact test scores. Insecurity about feeling different may harm the child's self-confidence, which can directly impact confidence at school. However, parents may adjust their parental investments to help their child cope with this insecurity (e.g. providing any type of mental health investments), which may come at a cost of educational investments (Yi et al., 2015). Unfortunately, our data does not allow us to distinguish confidence from asymmetric parental investments, and this is left for future research.

## 7 Conclusion

Gender gaps in educational performance are typically explained by gender-biased environments and socialization; the literature has paid little attention to the potential role of biology in creating these gender differences. This paper is the first to examine the role of biology as an additionally important factor and

[36] To identify children who are living in the biblebelt, we match the child's municipality of residence at the time of taking the test to the share of votes to the conservative Christian party (SGP) during the 2017 Dutch national elections in that same municipality. A municipality is defined as 'more religious' if the vote share for this particular party exceeds 1% (which holds for about 29% of the municipalities in our sample). We cannot match the child to the share of SGP votes for 114 children (0.2%) of the sample, which leaves 50,839 observations in our sample, and 42,961 in the full control specification.

[37] See column 5 and 6 of Table A13.

[38] The full estimation results are available in Table A13.

Table 11: By traditional gender norms

| | Aggregate score | | Reading score | | Math score | |
|---|---|---|---|---|---|---|
| | $DD_{male}$ | $DD_{female}$ | $DD_{male}$ | $DD_{female}$ | $DD_{male}$ | $DD_{female}$ |
| Baseline | -0.047 | -0.051* | -0.040 | -0.025 | -0.046 | -0.075** |
| | (0.030) | (0.030) | (0.032) | (0.031) | (0.031) | (0.032) |
| $N$ | $43,069$ | $43,069$ | $43,069$ | $43,069$ | $43,069$ | $43,069$ |
| Less religious | -0.056 | -0.040 | -0.053 | 0.000 | -0.044 | -0.065* |
| | (0.036) | (0.036) | (0.038) | (0.036) | (0.036) | (0.038) |
| $N$ | 30,504 | 30,504 | 30,504 | 30,504 | 30,504 | 30,504 |
| More religious | -0.032 | -0.089 | -0.019 | -0.091 | -0.058 | -0.115* |
| | (0.056) | (0.057) | (0.059) | (0.058) | (0.056) | (0.061) |
| $N$ | 12,457 | 12,457 | 12,457 | 12,457 | 12,457 | 12,457 |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

it specifically focuses on the role of prenatal testosterone in explaining gender differences in performance in 8th grade of primary school. Prenatal testosterone is not only responsible for the sexual differentiation of the male fetus, but is also said to wire the brain with masculine behavioral patterns. Since male-typical cognitive skills are related to boys' advantage in math, biological factors may well explain some part of the gender gap in math and reading test scores. If there is such a role for biology, the role of any discriminatory or gender-biased social factors is currently being overstated.

Boys are exposed to elevated levels of prenatal testosterone between the eighth and twenty-fourth week of gestation. Based on evidence from the biological literature for other mammals it is hypothesized that also in humans prenatal testosterone can transfer in-utero from the male twin to his uterus mate. This would imply that individuals with a male co-twin are exposed to higher levels of prenatal testosterone than individuals with a female co-twin. We argue that opposite-sex twinning can be exploited as a natural experiment generating quasi-experimental variation in prenatal testosterone exposure to test the link between prenatal testosterone and primary school test scores.

Using Dutch administrative data on all twins and a control group of closely spaced singletons (CSS), we find that girls with an opposite-sex twin sibling score 7% of a standard deviation lower on math, with no effects found on an aggregate and reading score. If opposite-sex twinning is indeed a good proxy for exposure to prenatal testosterone, these findings suggest that more prenatal testosterone leads to lower math test scores for girls. This result is rather counterintuitive as one would expect improved math performance for girls with increased exposure to prenatal testosterone. A series of robustness and sensitivity analysis shows that this effect is concentrated among children who are raised in families with more traditional gender norms. Possibly, children in these families are more sensitive to adhering to a social norm than children who grow up in less traditional families. A feeling of being different may result in adverse behavioral responses or behavior, which may divert the attention of the child and his parents away from performing well in school.

Our findings imply that biological factors seem to play a role in children's educational performance, but that these effects materialize depending on environmental factors. If these effects influence the type of education a child enrols for as adolescents, they may translate into different economic outcomes in adulthood, such as wage differences as was found earlier in Gielen et al. (2016). As such, a better understanding of the role of biological factors in generating gender differences in economic outcomes is crucial for the discussion on (the presence of) labor market discrimination and the required measures to limit it.

# References

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

Altonji, J. G. (1995). The effects of high school curriculum on education and labor market outcomes. *Journal of Human Resources*, 30(3):409–38.

Altonji, J. G., Blom, E., and Meghir, C. (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annu. Rev. Econ.*, 4(1):185–223.

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.

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, 121(1-2):343–375.

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.

Autor, D., Figlio, D., Karbownik, K., Roth, J., and Wasserman, M. (2017). Family disadvantage and the gender gap in behavioral and educational outcomes. *National Bureau of Economic Research.*, Working paper nr. 22267.

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.

Bertrand, M. and Pan, J. (2013). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1):32–64.

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.*

Bharadwaj, P., Eberhard, J. P., and Neilson, C. A. (2018). Health at birth, parental investments, and academic outcomes. *Journal of Labor Economics*, 36(2):349–394.

Björklund, A. and Jäntti, M. (2012). How important is family background for labor-economic outcomes? *Labour Economics*, 19(4):465–474.

Björklund, A., Lindahl, M., and Plug, E. (2006). The origins of intergenerational associations: Lessons from swedish adoption data. *The Quarterly Journal of Economics*, 121(3):999–1028.

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. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Blau, F. D., Kahn, L. M., Brummund, P., Cook, J., and Larson-Koester, M. (2017). Is there still son preference in the united states? *National Bureau of Economic Research. Working Paper No. 23816.*

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

Bouckaert, A., Theunissen, I., and Van, M. L. (1992). Weight and length of newborns. differences between boys and girls. *Journal de gynecologie, obstetrique et biologie de la reproduction*, 21(4):398–402.

Brenøe, A. A. (2018). Origins of gender norms: Sibling gender composition and womens choice of occupation and partner. *University of Zurich, Department of Economics, Working Paper No. 294.*

Brenøe, A. A. and Lundberg, S. (2017). Gender gaps in the effects of childhood family environment: Do they persist into adulthood? *European Economic Review.*

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.

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.

Chorny, V., Webbink, D., et al. (2010). The effect of accountability policies in primary education in amsterdam. *CPB Discussion Paper no. 144.*

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.

Cools, A. and Patacchini, E. (2017). Sibling gender composition and women's wages. *IZA Discussion Paper No. 11001.*

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.

Dahl, G. B. and Moretti, E. (2008). The demand for sons. *The Review of Economic Studies*, 75(4):1085–1120.

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.

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.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119.

Goldin, C., Katz, L. F., and Kuziemko, I. (2006). The homecoming of american college women: The reversal of the college gender gap. Technical report, National Bureau of Economic Research.

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 childbearing 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.

Joensen, J. S. and Nielsen, H. S. (2009). Is there a causal effect of high school math on labor market outcomes? *Journal of Human Resources*, 44(1):171–198.

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

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.

Lippmann, Q. and Senik, C. (2018). Math, girls and socialism. *Journal of Comparative Economics*, 46(3):874–888.

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.

Peter, N., Lundborg, P., Mikkelsen, S., and Webbink, D. (2018). The effect of a sibling's gender on siblings and family formation. *Labour Economics*, 54:61–78.

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.

Price, J. (2008). Parent-child quality time does birth order matter? *Journal of Human Resources*, 43(1):240–265.

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.

Reardon, S., Fahle, E., Kalogrides, D., Podolsky, A., and Zarate, R. (2018). Gender achievement gaps in u.s. school districts. *CEPA Working Paper No. 18-13.*

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

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.

Torun, H. and Tumen, S. (2016). The empirical content of season-of-birth effects: An investigation with turkish data. *IZA Discussion Paper*.

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.

Voldner, N., Frey Frøslie, K., Godang, K., Bollerslev, J., and Henriksen, T. (2009). Determinants of birth weight in boys and girls. *human_ontogenetics*, 3(1):7–12.

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, intra-household resource allocation and child outcomes. *The Economic Journal*, 125(588).

# Appendix

Table A1: Variable descriptions

| Variable | Dataset | Definition | Units |
|---|---|---|---|
| *A. Educational outcomes* | | | |
| Total score | CITO | Final aggregate Cito-score | Standardized |
| Language score | CITO | Cito language score | Standardized |
| Math score | CITO | Cito math score | Standardized |
| School advice | CITO | Teacher advice for a track in secondary education. Hierarchical with one representing the lowest school advice (lower vocational education) and nine representing the highest track (pre-university education). | Categorial, 1 to 9 |
| *B. Demographic variables* | | | |
| Age | GBA | Age at taking the test | Months |
| Parity | GBA | Birth order | Categorical |
| Spacing | GBA | Difference between sequential births | Months |
| Nonnative | GBA | Non-Dutch indicator | 0/1 dummy |
| Family size | GBA | Family size (number of siblings plus one) via mother. | |
| Mother's age | GBA | Maternal age at birth | Months. |
| Father's age | GBA | Paternal age at birth | Months. |
| *C. Household characteristics* | | | |
| HH-type | Huishoudens | Household type at the time of taking the test. | Categorical |
| HH-income | Baanpersjaartab & Zelfstandigentab | Household income in the year the child has its fourth birthday (combined income of parents from labor income and self-employment) | Euros/year |
| Mother working | Baanpersjaartab & Zelfstandigentab | Mother has positive earnings in the year the child has its fourth birthday | 0/1 dummy |
| Mother in DI | AOTOPERSOON-BUS | Mother has positive DI benefits in the year the child has its fourth | 0/1 dummy |
| *D. Birth outcomes* | | | |
| Birth weight | PRNL | Raw birth weight. | Grams |
| Gestation | PRNL | Gestational length. | Days |

Table A2: Results for aggregate test score (scale: 501-550) - by gender

| *A. Males* | | | | | |
|---|---|---|---|---|---|
| | Aggregate score | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Twin | 0.226*** | -0.002 | 0.220*** | 0.003 | -0.001 |
| | (0.025) | (0.021) | (0.026) | (0.023) | (0.023) |
| OS | 0.001 | 0.011 | -0.006 | 0.004 | 0.005 |
| | (0.030) | (0.025) | (0.032) | (0.027) | (0.027) |
| Twin*OS | 0.004 | -0.040 | -0.020 | -0.047 | -0.047 |
| | (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 |

| *B. Females* | | | | | |
|---|---|---|---|---|---|
| | Aggregate score | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Twin | 0.197*** | -0.051** | 0.180*** | -0.045* | -0.051** |
| | (0.025) | (0.022) | (0.027) | (0.023) | (0.023) |
| OS | 0.015 | -0.016 | 0.009 | -0.011 | -0.010 |
| | (0.030) | (0.025) | (0.032) | (0.027) | (0.027) |
| Twin*OS | -0.048 | -0.046* | -0.053 | -0.054* | -0.054* |
| | (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 |

$^{\text{Notes}}$ : Results are based on OLS. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Results for standardized reading and math scores - by gender

*A. Males*

| | Reading score | | | | | Math score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (1) | (2) | (3) | (4) | (5) |
| Twin | 0.273*** | 0.021 | 0.267*** | 0.022 | 0.018 | 0.112*** | -0.029 | 0.114*** | -0.019 | -0.021 |
| | (0.026) | (0.022) | (0.028) | (0.024) | (0.024) | (0.023) | (0.021) | (0.025) | (0.023) | (0.023) |
| OS | -0.005 | 0.003 | -0.004 | 0.005 | 0.006 | -0.003 | 0.006 | -0.016 | -0.006 | -0.005 |
| | (0.031) | (0.026) | (0.033) | (0.028) | (0.028) | (0.028) | (0.024) | (0.030) | (0.027) | (0.027) |
| Twin*OS | 0.012 | -0.032 | -0.014 | -0.041 | -0.041 | -0.008 | -0.041 | -0.026 | -0.045 | -0.046 |
| | (0.035) | (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 |

*B. Females*

| | Reading score | | | | | Math score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (1) | (2) | (3) | (4) | (5) |
| Twin | 0.212*** | -0.044** | 0.194*** | -0.046* | -0.051** | 0.126*** | -0.045** | 0.122*** | -0.031 | -0.035 |
| | (0.025) | (0.022) | (0.027) | (0.024) | (0.024) | (0.025) | (0.023) | (0.027) | (0.025) | (0.025) |
| OS | 0.019 | -0.011 | 0.011 | -0.012 | -0.011 | 0.005 | -0.015 | 0.008 | -0.002 | -0.002 |
| | (0.029) | (0.025) | (0.032) | (0.027) | (0.027) | (0.029) | (0.026) | (0.032) | (0.028) | (0.028) |
| Twin*OS | -0.028 | -0.024 | -0.030 | -0.027 | -0.027 | -0.063* | -0.064** | -0.075** | -0.079** | -0.079** |
| | (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 |

$^{\text{Notes}}$: Results are based on OLS. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Robustness results for aggregate test score (standardized) - sub-sample estimates

| | Aggregate score | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Baseline | First born | Native kids | Two-child families |
| Twin | -0.007 | 0.016 | -0.016 | 0.017 |
| | (0.023) | (0.036) | (0.026) | (0.044) |
| OS | 0.005 | -0.016 | -0.005 | 0.015 |
| | (0.027) | (0.046) | (0.033) | (0.050) |
| Female | -0.067** | -0.042 | -0.088** | -0.013 |
| | (0.028) | (0.045) | (0.035) | (0.054) |
| Twin*Female | -0.039 | -0.054 | -0.018 | -0.067 |
| | (0.031) | (0.049) | (0.038) | (0.059) |
| OS*Female | -0.018 | -0.001 | 0.034 | 0.020 |
| | (0.036) | (0.063) | (0.044) | (0.066) |
| Twin*OS | -0.047 | -0.033 | -0.016 | -0.059 |
| | (0.030) | (0.050) | (0.036) | (0.055) |
| Twin*OS*Female | -0.005 | -0.025 | -0.074 | -0.065 |
| | (0.040) | (0.068) | (0.049) | (0.073) |
| $DD_{male}$ | -0.047 | -0.033 | -0.016 | -0.059 |
| | (0.030) | (0.050) | (0.036) | (0.055) |
| $DD_{female}$ | -0.051* | -0.057 | -0.090** | -0.123** |
| | (0.030) | (0.048) | (0.037) | (0.055) |
| N | 43,069 | 19,576 | 34,003 | 14,034 |
| Controls | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Robustness results for standardized reading and math score - sub-sample estimates

| | Reading score | | | | Math score | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | Baseline | First born | Native kids | Two-child families | Baseline | First born | Native kids | Two-child families |
| Twin | 0.022 | 0.047 | 0.007 | 0.058 | -0.036 | -0.011 | -0.034 | -0.019 |
| | (0.023) | (0.037) | (0.027) | (0.045) | (0.023) | (0.037) | (0.026) | (0.046) |
| OS | 0.006 | -0.025 | -0.006 | -0.009 | -0.006 | -0.002 | -0.013 | 0.019 |
| | (0.028) | (0.048) | (0.035) | (0.051) | (0.027) | (0.046) | (0.032) | (0.050) |
| Female | 0.210*** | 0.210*** | 0.186*** | 0.262*** | -0.392*** | -0.332*** | -0.407*** | -0.363*** |
| | (0.029) | (0.046) | (0.036) | (0.054) | (0.029) | (0.049) | (0.036) | (0.058) |
| Twin*Female | -0.077** | -0.075 | -0.058 | -0.116** | 0.014 | -0.034 | 0.034 | 0.011 |
| | (0.032) | (0.050) | (0.039) | (0.059) | (0.032) | (0.053) | (0.039) | (0.063) |
| OS*Female | -0.020 | 0.002 | 0.035 | 0.012 | 0.002 | -0.007 | 0.048 | 0.069 |
| | (0.037) | (0.065) | (0.046) | (0.067) | (0.037) | (0.066) | (0.046) | (0.070) |
| Twin*OS | -0.040 | -0.022 | -0.014 | -0.035 | -0.046 | -0.051 | -0.015 | -0.070 |
| | (0.032) | (0.052) | (0.038) | (0.057) | (0.031) | (0.050) | (0.036) | (0.056) |
| Twin*OS*Female | 0.015 | -0.014 | -0.049 | -0.038 | -0.029 | -0.017 | -0.097* | -0.113 |
| | (0.042) | (0.070) | (0.050) | (0.075) | (0.042) | (0.072) | (0.051) | (0.077) |
| $DD_{male}$ | -0.040 | -0.022 | -0.014 | -0.035 | -0.046 | -0.051 | -0.015 | -0.070 |
| | (0.032) | (0.052) | (0.038) | (0.057) | (0.031) | (0.050) | (0.036) | (0.056) |
| $DD_{female}$ | -0.025 | -0.036 | -0.064* | -0.073 | -0.075** | -0.067 | -0.112*** | -0.182*** |
| | (0.031) | (0.048) | (0.037) | (0.055) | (0.032) | (0.052) | (0.039) | (0.059) |
| N | 43,069 | 19,576 | 34,003 | 14,034 | 43,069 | 19,576 | 34,003 | 14,034 |
| Controls | Y | Y | Y | Y | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Robustness results for aggregate test score (standardized) - different age bandwidth for CSS

| | Aggregate score (std) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | 12 months | 18 months | 24 months | 36 months |
| Twin | -0.007 | -0.072*** | -0.082*** | -0.055*** |
| | (0.023) | (0.012) | (0.011) | (0.010) |
| OS | 0.005 | 0.019** | 0.021*** | 0.019*** |
| | (0.027) | (0.008) | (0.005) | (0.004) |
| Female | -0.067** | -0.048*** | -0.055*** | -0.053*** |
| | (0.028) | (0.008) | (0.005) | (0.004) |
| Twin*Female | -0.039 | -0.060*** | -0.054*** | -0.056*** |
| | (0.031) | (0.016) | (0.015) | (0.014) |
| OS*Female | -0.018 | -0.036*** | -0.035*** | -0.030*** |
| | (0.036) | (0.011) | (0.007) | (0.005) |
| Twin*OS | -0.047 | -0.063*** | -0.065*** | -0.064*** |
| | (0.030) | (0.017) | (0.015) | (0.015) |
| Twin*OS*Female | -0.005 | 0.016 | 0.015 | 0.010 |
| | (0.040) | (0.022) | (0.020) | (0.020) |
| $DD_{male}$ | -0.047 | -0.063*** | -0.065*** | -0.064*** |
| | (0.030) | (0.017) | (0.015) | (0.015) |
| $DD_{female}$ | -0.051* | -0.047*** | -0.050*** | -0.054*** |
| | (0.030) | (0.016) | (0.015) | (0.015) |
| N | 43,069 | 132,650 | 279,980 | 492,264 |
| Controls | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Robustness results for standardized reading and math score - different age bandwidth for CSS

| | Language score (std) | | | | Math score (std) | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | 12 months | 18 months | 24 months | 36 months | 12 months | 18 months | 24 months | 36 months |
| Twin | 0.022 | -0.049*** | -0.065*** | -0.046*** | -0.036 | -0.081*** | -0.080*** | -0.046*** |
| | (0.023) | (0.012) | (0.011) | (0.011) | (0.023) | (0.012) | (0.011) | (0.010) |
| OS | 0.006 | 0.033*** | 0.033*** | 0.032*** | -0.006 | -0.009 | -0.004 | -0.005 |
| | (0.028) | (0.008) | (0.005) | (0.004) | (0.027) | (0.008) | (0.005) | (0.004) |
| Female | 0.210*** | 0.201*** | 0.190*** | 0.197*** | -0.392*** | -0.344*** | -0.345*** | -0.346*** |
| | (0.029) | (0.009) | (0.005) | (0.004) | (0.029) | (0.009) | (0.005) | (0.004) |
| Twin*Female | -0.077** | -0.070*** | -0.060*** | -0.067*** | 0.014 | -0.034** | -0.035** | -0.033** |
| | (0.032) | (0.017) | (0.015) | (0.015) | (0.032) | (0.017) | (0.015) | (0.015) |
| OS*Female | -0.020 | -0.048*** | -0.042*** | -0.044*** | 0.002 | -0.010 | -0.015** | -0.005 |
| | (0.037) | (0.011) | (0.007) | (0.005) | (0.037) | (0.011) | (0.007) | (0.005) |
| Twin*OS | -0.040 | -0.069*** | -0.070*** | -0.069*** | -0.046 | -0.046*** | -0.052*** | -0.050*** |
| | (0.032) | (0.017) | (0.016) | (0.016) | (0.031) | (0.017) | (0.016) | (0.015) |
| Twin*OS*Female | 0.015 | 0.045** | 0.040** | 0.041** | -0.029 | -0.015 | -0.010 | -0.020 |
| | (0.042) | (0.022) | (0.020) | (0.020) | (0.042) | (0.023) | (0.021) | (0.021) |
| $DD_{male}$ | -0.040 | -0.069*** | -0.070*** | -0.069*** | -0.046 | -0.046*** | -0.052*** | -0.050*** |
| | (0.032) | (0.017) | (0.016) | (0.016) | (0.031) | (0.017) | (0.016) | (0.015) |
| $DD_{female}$ | -0.025 | -0.024 | -0.030** | -0.029* | -0.075** | -0.061*** | -0.062*** | -0.070*** |
| | (0.031) | (0.016) | (0.015) | (0.015) | (0.032) | (0.018) | (0.016) | (0.016) |
| N | 43,069 | 132,650 | 279,980 | 492,264 | 43,069 | 132,650 | 279,980 | 492,264 |
| Controls | Y | Y | Y | Y | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Robustness: matching estimators

| | Aggregate score | | Reading score | | Math score | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Twin | 0.018 | -0.030 | 0.042 | -0.004 | -0.003 | -0.056** |
| | (0.027) | (0.025) | (0.027) | (0.027) | (0.027) | (0.025) |
| OS | 0.014 | 0.000 | 0.016 | 0.005 | 0.002 | -0.018 |
| | (0.034) | (0.030) | (0.036) | (0.033) | (0.033) | (0.029) |
| Female | -0.051 | -0.083*** | 0.211*** | 0.188*** | -0.355*** | -0.400*** |
| | (0.036) | (0.032) | (0.037) | (0.033) | (0.037) | (0.032) |
| Twin*Female | -0.058 | -0.023 | -0.082** | -0.043 | -0.026 | 0.016 |
| | (0.039) | (0.036) | (0.039) | (0.037) | (0.040) | (0.037) |
| OS*Female | 0.006 | -0.005 | 0.001 | -0.010 | 0.021 | 0.024 |
| | (0.046) | (0.041) | (0.048) | (0.044) | (0.048) | (0.042) |
| Twin*OS | -0.056 | -0.039 | -0.049 | -0.030 | -0.054 | -0.034 |
| | (0.037) | (0.035) | (0.039) | (0.038) | (0.037) | (0.035) |
| Twin*OS*Female | -0.026 | -0.028 | -0.003 | -0.007 | -0.045 | -0.064 |
| | (0.050) | (0.048) | (0.051) | (0.050) | (0.052) | (0.049) |
| $DD_{male}$ | -0.056 | -0.039 | -0.049 | -0.030 | -0.054 | -0.034 |
| | (0.037) | (0.035) | (0.039) | (0.038) | (0.037) | (0.035) |
| $DD_{female}$ | -0.081** | -0.066* | -0.052 | -0.038 | -0.099** | -0.098** |
| | (0.037) | (0.036) | (0.038) | (0.036) | (0.040) | (0.038) |
| N | 43,068 | 33,029 | 43,068 | 33,029 | 43,068 | 33,029 |
| Kernel M | Y | N | Y | N | Y | N |
| Inverse Prob. | N | Y | N | Y | N | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. 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. Kernel matching based on Epanechnikov kernel with a bandwidth of 0.06). The results of inverse probability matching excludes observations with propensity scores lower than 0.1 and higher than 0.9. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Birth outcomes

| | Birth weight (grams) | Birth weight (grams) | Gestation (in days) | Gestation (in days) | Gestation (in weeks) | Gestation (in weeks) |
|---|---|---|---|---|---|---|
| Twin | -826.690*** | -807.474*** | -18.594*** | -18.016*** | -2.638*** | -2.552*** |
| | (13.588) | (13.624) | (0.398) | (0.406) | (0.057) | (0.058) |
| OS | 28.970* | 29.118* | 0.566 | 0.547 | 0.083 | 0.080 |
| | (16.401) | (16.157) | (0.466) | (0.467) | (0.067) | (0.067) |
| Female | -75.761*** | -77.025*** | 2.066*** | 1.977*** | 0.306*** | 0.293*** |
| | (17.617) | (17.461) | (0.492) | (0.494) | (0.070) | (0.071) |
| Twin*Female | -9.447 | -10.567 | -1.272** | -1.227** | -0.192** | -0.186** |
| | (19.259) | (19.007) | (0.564) | (0.563) | (0.081) | (0.081) |
| OS*Female | -41.317* | -39.087* | -0.992 | -0.880 | -0.159* | -0.143 |
| | (21.184) | (20.953) | (0.612) | (0.616) | (0.088) | (0.088) |
| Twin*OS | 66.274*** | 64.586*** | 2.375*** | 2.324*** | 0.343*** | 0.337*** |
| | (18.235) | (17.903) | (0.537) | (0.535) | (0.077) | (0.077) |
| Twin*OS*Female | 5.252 | 5.522 | 0.166 | 0.104 | 0.041 | 0.032 |
| | (22.932) | (22.625) | (0.671) | (0.673) | (0.096) | (0.097) |
| $DD_{males}$ | 66.274*** | 64.586*** | 2.375*** | 2.324*** | 0.343*** | 0.337*** |
| | (18.235) | (17.903) | (0.537) | (0.535) | (0.077) | (0.077) |
| $DD_{females}$ | 71.525*** | 70.108*** | 2.541*** | 2.428*** | 0.384*** | 0.368*** |
| | (17.814) | (17.607) | (0.527) | (0.527) | (0.076) | (0.076) |
| N | $80,663$ | $80,663$ | $80,663$ | $80,663$ | $80,663$ | $80,663$ |
| Controls | N | Y | N | Y | N | Y |

$^{\text{Notes:}}$ Results are based on OLS model. Controls are birth order dummies, maternal age at birth, non-native dummy, and year of birth dummies. Standard errors are clustered on maternal ID and are in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Mechanisms at work - standardized aggregate score

| | Aggregate score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | Baseline | Two parent HH | One parent HH | Non-divorced | Divorced | Not married | Low-income | High-income | Dis-advantaged | Advantaged |
| Twin | -0.007 | 0.005 | -0.071 | 0.018 | -0.177*** | 0.028 | -0.035 | 0.039 | -0.059 | -0.015 |
| | (0.023) | (0.025) | (0.055) | (0.026) | (0.060) | (0.068) | (0.028) | (0.036) | (0.055) | (0.026) |
| OS | 0.005 | 0.028 | -0.109* | 0.036 | -0.132* | -0.033 | -0.001 | 0.020 | -0.036 | -0.004 |
| | (0.027) | (0.030) | (0.063) | (0.031) | (0.071) | (0.082) | (0.032) | (0.046) | (0.057) | (0.032) |
| Female | -0.067** | -0.072** | -0.063 | -0.080** | -0.098 | 0.059 | -0.055* | -0.096* | -0.086 | -0.108*** |
| | (0.028) | (0.031) | (0.066) | (0.033) | (0.073) | (0.080) | (0.033) | (0.049) | (0.058) | (0.034) |
| Twin*Female | -0.039 | -0.047 | 0.030 | -0.039 | 0.067 | -0.156* | -0.032 | -0.032 | -0.003 | -0.004 |
| | (0.031) | (0.035) | (0.076) | (0.036) | (0.084) | (0.091) | (0.039) | (0.053) | (0.074) | (0.038) |
| OS*Female | -0.018 | -0.019 | 0.017 | -0.001 | -0.046 | -0.106 | -0.051 | 0.073 | -0.005 | 0.026 |
| | (0.036) | (0.040) | (0.085) | (0.041) | (0.095) | (0.107) | (0.043) | (0.063) | (0.076) | (0.044) |
| Twin*OS | -0.047 | -0.070** | 0.068 | -0.083** | 0.169** | -0.062 | -0.010 | -0.095* | -0.010 | -0.028 |
| | (0.030) | (0.033) | (0.076) | (0.035) | (0.084) | (0.095) | (0.038) | (0.050) | (0.074) | (0.036) |
| Twin*OS*Female | -0.005 | 0.004 | -0.079 | -0.011 | -0.080 | 0.133 | -0.019 | -0.039 | -0.003 | -0.052 |
| | (0.040) | (0.044) | (0.101) | (0.046) | (0.111) | (0.123) | (0.050) | (0.068) | (0.097) | (0.049) |
| $DD_{male}$ | -0.047 | -0.070** | 0.068 | -0.083** | 0.169** | -0.062 | -0.010 | -0.095* | -0.010 | -0.028 |
| | (0.030) | (0.033) | (0.076) | (0.035) | (0.084) | (0.095) | (0.038) | (0.050) | (0.074) | (0.036) |
| $DD_{female}$ | -0.051* | -0.066** | -0.011 | -0.094*** | 0.089 | 0.071 | -0.030 | -0.133** | -0.013 | -0.080** |
| | (0.030) | (0.033) | (0.075) | (0.035) | (0.083) | (0.091) | (0.037) | (0.052) | (0.069) | (0.037) |
| N | 43,069 | 36,404 | 6,383 | 33,374 | 5,370 | 4,325 | 25,429 | 17,640 | 6,352 | 30,647 |
| Controls | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Mechanisms a work - standardized reading score

| | Reading score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | Baseline | Two parent HH | One parent HH | Non-divorced | Divorced | Not married | Low-income | High-income | Dis-advantaged | Advantaged |
| Twin | 0.022 | 0.028 | -0.027 | 0.035 | -0.105 | 0.076 | -0.000 | 0.038 | 0.038 | -0.005 |
| | (0.023) | (0.026) | (0.055) | (0.027) | (0.064) | (0.069) | (0.030) | (0.037) | (0.057) | (0.028) |
| OS | 0.006 | 0.020 | -0.074 | 0.025 | -0.108 | 0.006 | 0.016 | -0.017 | 0.027 | -0.016 |
| | (0.028) | (0.032) | (0.064) | (0.033) | (0.077) | (0.084) | (0.034) | (0.048) | (0.059) | (0.034) |
| Female | 0.210*** | 0.186*** | 0.290*** | 0.174*** | 0.273*** | 0.355*** | 0.252*** | 0.109** | 0.262*** | 0.149*** |
| | (0.029) | (0.032) | (0.067) | (0.034) | (0.077) | (0.083) | (0.035) | (0.051) | (0.061) | (0.035) |
| Twin*Female | -0.077** | -0.068* | -0.067 | -0.055 | -0.076 | -0.193** | -0.087** | -0.013 | -0.137* | -0.014 |
| | (0.032) | (0.036) | (0.077) | (0.037) | (0.087) | (0.094) | (0.040) | (0.055) | (0.076) | (0.039) |
| OS*Female | -0.020 | 0.001 | -0.074 | 0.014 | -0.103 | -0.127 | -0.072 | 0.117* | -0.077 | 0.037 |
| | (0.037) | (0.041) | (0.088) | (0.043) | (0.100) | (0.113) | (0.045) | (0.065) | (0.079) | (0.046) |
| Twin*OS | -0.040 | -0.060* | 0.076 | -0.064* | 0.131 | -0.078 | -0.012 | -0.058 | -0.077 | -0.006 |
| | (0.032) | (0.035) | (0.077) | (0.037) | (0.090) | (0.097) | (0.040) | (0.052) | (0.076) | (0.038) |
| Twin*OS*Female | 0.015 | 0.000 | 0.039 | -0.014 | 0.064 | 0.132 | 0.021 | -0.067 | 0.117 | -0.053 |
| | (0.042) | (0.046) | (0.103) | (0.048) | (0.116) | (0.128) | (0.052) | (0.071) | (0.100) | (0.051) |
| $DD_{male}$ | -0.040 | -0.060* | 0.076 | -0.064* | 0.131 | -0.078 | -0.012 | -0.058 | -0.077 | -0.006 |
| | (0.032) | (0.035) | (0.077) | (0.037) | (0.090) | (0.097) | (0.040) | (0.052) | (0.076) | (0.038) |
| $DD_{female}$ | -0.025 | -0.059* | 0.115 | -0.078** | 0.194** | 0.055 | 0.009 | -0.125** | 0.040 | -0.059 |
| | (0.031) | (0.034) | (0.075) | (0.035) | (0.083) | (0.092) | (0.037) | (0.053) | (0.070) | (0.037) |
| N | 43,069 | 36,404 | 6,383 | 33,374 | 5,370 | 4,325 | 25,429 | 17,640 | 6,352 | 30,647 |
| Controls | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Mechanisms at work - standardized math score

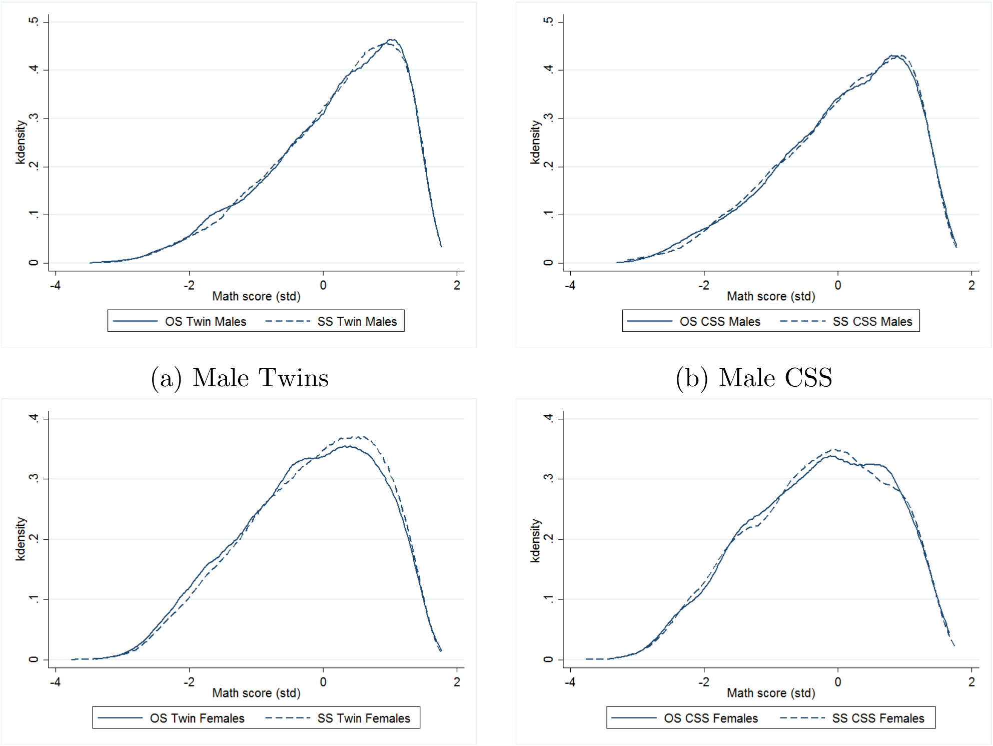
| | Aggregate score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | Baseline | Two parent HH | One parent HH | Non-divorced | Divorced | Not married | Low-income | High-income | Dis-advantaged | Advantaged |
| Twin | -0.036 | -0.022 | -0.097* | 0.001 | -0.225*** | -0.046 | -0.065** | 0.040 | -0.152*** | -0.025 |
| | (0.023) | (0.025) | (0.057) | (0.026) | (0.061) | (0.072) | (0.028) | (0.037) | (0.056) | (0.027) |
| OS | -0.006 | 0.017 | -0.115* | 0.028 | -0.113 | -0.085 | -0.033 | 0.059 | -0.110* | 0.005 |
| | (0.027) | (0.029) | (0.065) | (0.030) | (0.071) | (0.086) | (0.032) | (0.045) | (0.056) | (0.032) |
| Female | -0.392*** | -0.381*** | -0.442*** | -0.379*** | -0.513*** | -0.303*** | -0.417*** | -0.330*** | -0.470*** | -0.415*** |
| | (0.029) | (0.032) | (0.068) | (0.034) | (0.076) | (0.084) | (0.034) | (0.052) | (0.059) | (0.035) |
| Twin*Female | 0.014 | -0.003 | 0.105 | -0.009 | 0.222** | -0.103 | 0.037 | -0.043 | 0.136* | 0.025 |
| | (0.032) | (0.036) | (0.079) | (0.037) | (0.087) | (0.096) | (0.040) | (0.056) | (0.076) | (0.039) |
| OS*Female | 0.002 | -0.014 | 0.095 | -0.002 | 0.022 | -0.037 | 0.004 | 0.009 | 0.079 | 0.024 |
| | (0.037) | (0.041) | (0.089) | (0.043) | (0.099) | (0.113) | (0.044) | (0.065) | (0.077) | (0.046) |
| Twin*OS | -0.046 | -0.063* | 0.020 | -0.085** | 0.150* | -0.032 | -0.005 | -0.128*** | 0.074 | -0.052 |
| | (0.031) | (0.033) | (0.078) | (0.035) | (0.083) | (0.098) | (0.039) | (0.049) | (0.075) | (0.036) |
| Twin*OS*Female | -0.029 | -0.006 | -0.154 | -0.010 | -0.208* | 0.092 | -0.070 | 0.012 | -0.120 | -0.050 |
| | (0.042) | (0.046) | (0.105) | (0.048) | (0.115) | (0.130) | (0.052) | (0.071) | (0.099) | (0.051) |
| $DD_{male}$ | -0.046 | -0.063* | 0.020 | -0.085** | 0.150* | -0.032 | -0.005 | -0.128*** | 0.074 | -0.052 |
| | (0.031) | (0.033) | (0.078) | (0.035) | (0.083) | (0.098) | (0.039) | (0.049) | (0.075) | (0.036) |
| $DD_{female}$ | -0.075** | -0.069* | -0.134* | -0.095** | -0.058 | 0.060 | -0.074* | -0.116** | -0.046 | -0.102** |
| | (0.032) | (0.035) | (0.080) | (0.037) | (0.088) | (0.097) | (0.039) | (0.056) | (0.072) | (0.039) |
| N | 43,069 | 36,404 | 6,383 | 33,374 | 5,370 | 4,325 | 25,429 | 17,640 | 6,352 | 30,647 |
| Controls | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: By traditional family norms

| | Aggregate score | | Reading score | | Math score | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | Less<br>Religious | More<br>Religious | Less<br>Religious | More<br>Religious | Less<br>Religious | More<br>Religious |
| Twin | -0.015 | 0.014 | 0.026 | 0.018 | -0.056** | 0.011 |
| | (0.027) | (0.041) | (0.028) | (0.042) | (0.027) | (0.042) |
| OS | 0.004 | 0.014 | 0.010 | 0.004 | -0.019 | 0.035 |
| | (0.032) | (0.049) | (0.034) | (0.052) | (0.032) | (0.049) |
| Female | -0.064* | -0.070 | 0.217*** | 0.205*** | -0.386*** | -0.405*** |
| | (0.033) | (0.052) | (0.035) | (0.054) | (0.034) | (0.055) |
| Twin*Female | -0.045 | -0.036 | -0.090** | -0.062 | 0.010 | 0.021 |
| | (0.037) | (0.058) | (0.038) | (0.060) | (0.038) | (0.061) |
| OS*Female | -0.034 | 0.026 | -0.050 | 0.049 | 0.001 | 0.014 |
| | (0.042) | (0.067) | (0.044) | (0.070) | (0.044) | (0.070) |
| Twin*OS | -0.056 | -0.032 | -0.053 | -0.019 | -0.044 | -0.058 |
| | (0.036) | (0.056) | (0.038) | (0.059) | (0.036) | (0.056) |
| Twin*OS*Female | 0.016 | -0.057 | 0.054 | -0.072 | -0.021 | -0.057 |
| | (0.048) | (0.076) | (0.050) | (0.079) | (0.050) | (0.079) |
| $DD_{male}$ | -0.056 | -0.032 | -0.053 | -0.019 | -0.044 | -0.058 |
| | (0.036) | (0.056) | (0.038) | (0.059) | (0.036) | (0.056) |
| $DD_{female}$ | -0.040 | -0.089 | 0.000 | -0.091 | -0.065* | -0.115* |
| | (0.036) | (0.057) | (0.036) | (0.058) | (0.038) | (0.061) |
| N | 30,504 | 12,457 | 30,504 | 12,457 | 30,504 | 12,457 |
| Controls | Y | Y | Y | Y | Y | Y |
| Income controls | Y | Y | Y | Y | Y | Y |

Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.



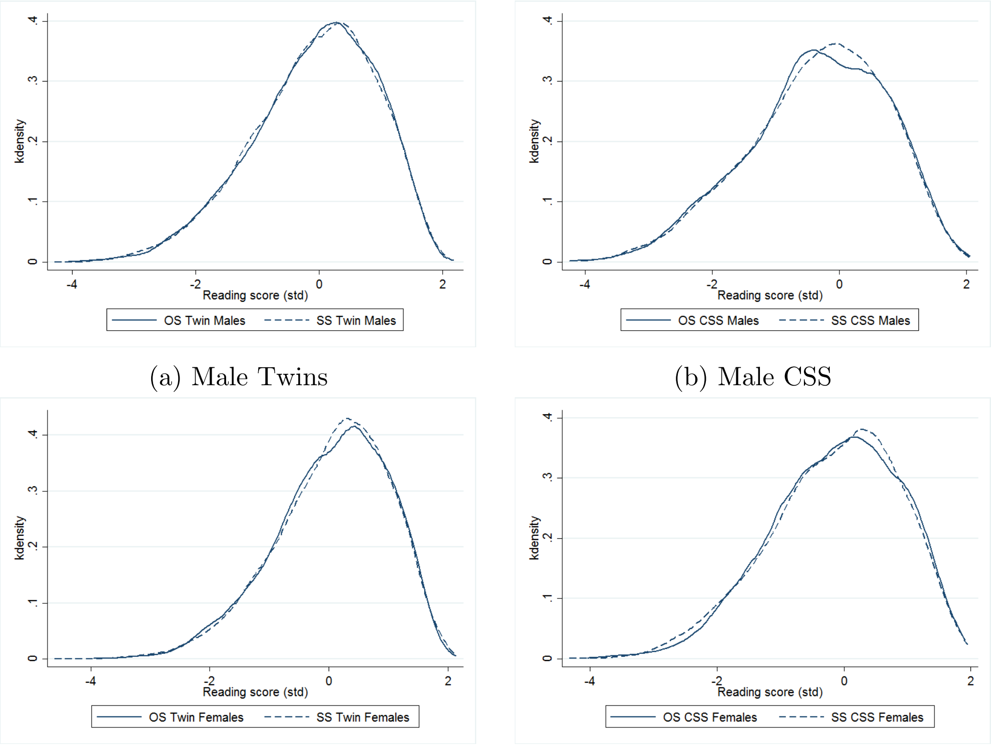
(a) Male Twins

(b) Male CSS

(c) Female Twins

(d) Female CSS

Figure A1: Test-score distributions for math, by gender, sibling-type and sibling gender.



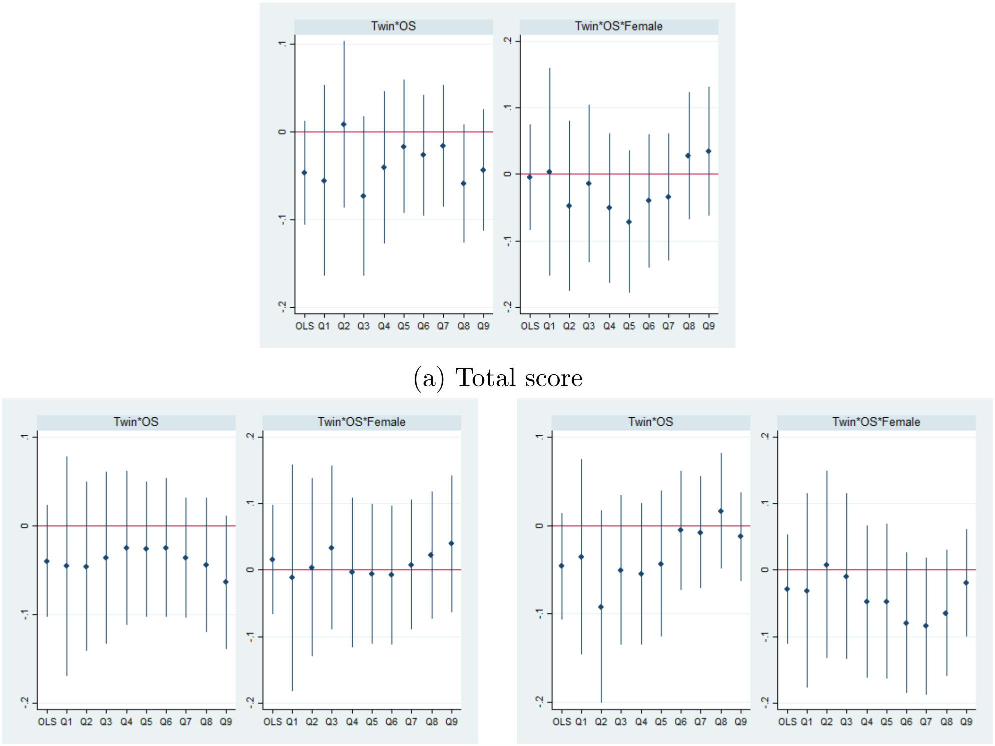
(a) Male Twins

(b) Male CSS

(c) Female Twins

(d) Female CSS

Figure A2: Test-score distributions for reading, by gender, sibling-type and sibling gender.



(a) Total score

(b) Reading score

(c) Math score

Figure A3: Quantile regression, and 95% confidence interval