About Family and Fate: Childhood Circumstances and Human Capital Formation

ISBN: 978 90 3610 559 0

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 740 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

About Family and Fate:
Childhood circumstances and human capital formation

Over familie en levenslot:
Omstandigheden in de kinderjaren en de vorming van menselijk kapitaal

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the rector magnificus

Prof. dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board.

The public defense shall be held on

Thursday, July 4, 2019 at 13:30 hours

by

Esmée Selina Zwiers

born in Gouda, the Netherlands

Erasmus University Rotterdam

Doctorate Committee:

Promotor:	Prof. dr. H.D. Webbink
Other members:	Dr. J.L.W. van Kippersluis
	Prof. dr. M. Lindeboom
	Prof. dr. B.J. ter Weel
Copromotor:	Dr. A.C. Gielen

Voor mijn ouders,
Bert en Hélène

Acknowledgements

I was fifteen when I decided that I wanted to study at Erasmus University Rotterdam. Although I first considered a degree in Business, I eventually decided that Economics would be a better fit. I believe that this proved to be a good decision. Now it has been almost nine years after starting my undergraduate studies, years with highs and lows, and I am very proud of the manuscript that lies in front of you.

I am grateful to many people that have supported me in my journey, but in particular I am thankful to my advisor: Dr. Anne Gielen. Anne, thank you for your guidance over the past years, I couldn't have wished for a better person to make me acquainted with academia. Most of all, thank you for getting me out of my comfort zone. I believe that this has not only contributed to my development as a researcher, but above all as a person. You are a mentor, but first and foremost a friend. I hope that our paths will cross many times in the future.

I want to thank my promotor Professor Dinand Webbink for his support and for always taking the time to comment on my work. Professor Olivier Marie, thank you for always providing me with suggestions on my work, for your support on the Job Market, and for being part of my committee. I am grateful to Prof. dr. Emilia del Bono, Prof. dr. Robert Dur, Dr. Hans van Kippersluis, Prof. dr. Maarten Lindeboom, and Prof. dr. Bas

ter Weel for reading my manuscript and taking place in the committee.

The Department of Economics has felt like home over the past years and I am grateful to all colleagues who have contributed to the great atmosphere. I am also grateful for all opportunities the department has offered me to develop myself as a researcher. The support staff of both the Department as well as the Tinbergen Institute provided excellent assistance throughout the years, and I especially want to thank Ankimon for her help. Martijn, thank you for giving me the opportunity to get familiar with academic research early in my undergraduate studies. Suzanne, thank you for supporting me from the time I wrote my undergraduate thesis with you. Finally, I want to thank my fellow Ph.D.s as this journey would have been less fun without them. In particular I thank Albert Jan, Anoek, Matthijs, and Megan for being great friends in joys and sorrows.

Spending three months at Princeton University has been an amazing experience, and I am excited to return very soon. I am very thankful to Professor Janet Currie for hosting my stay and for her support. Alice and Olexiy, your company made my stay extraordinary.

I also want to thank my family and friends for mental support and valuable distractions, and at times even actual research assistance (thanks Mom and Laura). I thank my grandmother and late grandfather for always believing in me. My friends outside academia - in particular Daan, Jeroen, Joyce, Jurriën, Josephine, Michelle, Renske and Sander - thank you for your companionship. Finally, I am very grateful for my paranymphs who will be by my side. Megan, ever since you joined the department we have been like two peas in a pod. Laura, you are a great little sister, and I can't wait to see what life has in store for you.

Last but definitely not least, I want to thank my parents: Bert and Hélène. Your unconditional love and support are the foundation of this manuscript. You taught me to work hard for the things I want, gave me perspective when I needed it, and always told me to have faith in myself. I believe that this manuscript is tangible evidence that your parental investments have paid off. You are the best!

Esmée Zwiers

Rotterdam, April 2019

Contents

Chapter 1

Introduction

The environment in which you grow up impacts the rest of your life. In the US, higher parental income is associated with higher child income in adulthood, higher college attendance, and lower teenage childbearing for women (Chetty et al., 2014). In the Netherlands, 28% of children end up in the bottom quintile of the income distribution if their parents were in the bottom quintile as well (Van der Brakel and Moonen, 2013). Parental income is just one way through which childhood circumstances may affect later life outcomes. However, many factors that could explain the relationship between childhood circumstances and later life outcomes - e.g. genetics, family stability, child health, the school the child attends, or the neighborhood the child grows up in - may be correlated with one another. This makes it hard to identify the contribution of each individual factor. It is for this reason that the focus within Economics lies on measuring *causal effects*: what would happen to child outcome Y in absence of childhood circumstance X?

Estimating causal effects is difficult as there likely are many unobserved

characteristics that correlate with childhood circumstance X. Economists often rely on natural experiments that allow for a separation of the effect of childhood circumstance X and these unobserved characteristics. These natural experiments can arise from policy interventions by which rules determine that some individuals are affected whereas others are not, from regional differences, or from nature. Evidence from this literature shows that characteristics early in life can explain a large part of the variation in educational attainment and labor market outcomes (Almond and Currie, 2011a; Currie, 2011). These characteristics range from differences in the prenatal environment (Almond and Currie, 2011b; Scholte et al., 2015; Almond et al., 2018), to differences in post-birth factors, like parental investments (Björklund and Salvanes, 2011), household budget constraints (Becker and Tomes, 1986; Cunha and Heckman, 2007), and parental enrollment in social insurance (Dahl and Gielen, 2018).

Childhood circumstances are important, which implies that inequalities that exist between children in early life may amplify in adulthood. Given that early childhood is a crucial time for skill formation (e.g. Heckman and Mosso, 2014), governments can allocate resources to earlier in the life-cycle to account for differences in childhood endowments. This is particularly important in a time of skill-biased technological change where increased returns to higher education may further reinforce existing disparities (Goldin and Katz, 2007; Acemoglu and Autor, 2011). This thesis adds to the literature by studying how three different childhood factors influence the child's human capital formation, and focuses on the role of family in fate.

Chapter two focuses on childhood circumstances *before conception*. This chapter studies how improved socioeconomic conditions affect parental se-

lection, i.e. which type of parents are having a child, and consequently the child's labor market and health outcomes in adulthood. I exploit novel variation that arose with the end of WWII in the Netherlands. Specifically, I use regional variation in the timing of the liberation, that caused a birth peak in the area that experienced the largest improvements in socioeconomic conditions.

I combine this regional variation with the timing of the liberation in a difference-in-difference framework. My main findings are twofold. First, I find no evidence that better socioeconomic conditions induced by the end of war lead to parental selection. Second, when focusing on a group of unplanned conceptions that occurred as a response to the improved circumstances, I find that growing up in a less stable family environment does not lead to poorer long-term labor market and health outcomes. Given the severity of the shock of the end of war, it is surprising that I find no evidence of parental selection in this chapter.

The third chapter examines childhood circumstances *after conception but before birth*. This chapter, which is joint work with Dr. Anne Gielen, examines whether biology, and specifically prenatal testosterone, can explain gender differences in educational performance. It is impossible to directly relate prenatal testosterone to educational outcomes, and therefore we exploit exogenous variation in prenatal testosterone exposure resulting from a natural experiment in twins, where prenatal testosterone is hypothesized to transfer in-utero from a male twin to his twin sibling. To control for the socialization effects of growing up with a same-sex or opposite-sex sibling we use a control group of closely spaced siblings whose birth dates are at most twelve months apart. We can identify the effect of prenatal

testosterone if socialization is similar for this group and twins.

We find that girls with a twin brother score about 7% of a standard deviation lower on math in primary school; no effects are found on reading. This seemingly counterintuitive effect is concentrated among children raised in traditional families and areas. We hypothesize that adherence to the social norm plays an important role. Our results are not just driven by biology, but rather materialize depending on environmental factors.

The fourth chapter focuses on childhood circumstances *after birth*. It studies how a preference against being in last place shapes the child's educational outcomes. These educational outcomes are determining the track level in secondary education, which is important given the effects of ability tracking that persist until adulthood. This chapter studies a Dutch educational reform that merged the two lowest educational tracks. Thus, those who would first enroll in the second-lowest track would now enroll in the lowest track. I exploit that only children who are expected to attend the lowest track are affected by the reform, and I analyze how the reform impacts the variables that determine track allocation in secondary school: (1) a track recommendation made by the teacher and (2) performance on a high-stakes standardized test.

The results show that children who are expected to attend the lowest track are less likely to receive a track recommendation above the lowest track, and perform worse on the high-stakes standardized cognitive test. The lower test score cannot entirely account for the lower track recommendation. Moreover, the effects are concentrated among the child's weakest subject (reading or math). Also, the effects are stronger for children in families where last place aversion is hypothesized to be stronger. This suggests

that children perform worse when the pressure to perform well is high, and hence last place aversion may affect performance negatively.

This dissertation examines how three different childhood factors affect human capital formation. It shows that better socioeconomic conditions induced by the end of WWII in the Netherlands do not lead to parental selection, that higher exposure to prenatal testosterone leads to lower math test scores for girls, and that last place aversion can cause lower educational performance. The universe of childhood circumstances that may influence the child's human capital formation - ranging from pre-conception factors to factors that appear long after birth - is large, and the three factors studied in this dissertation represent only small pieces of the puzzle of how one's family may influence one's fate. However, one can only solve a puzzle when all pieces are in the box, and hereby this dissertation brings us one step closer to understanding which childhood circumstances are important determinants of the child's human capital.

Chapter 2

The end of war, parental selection and children's outcomes[1]

2.1 Introduction

This paper explores how the socioeconomic environment at conception affects the size and composition of a cohort. Using a natural experiment induced by the timing of the end of war, I examine whether children who are conceived during a period of lower socioeconomic turbulence grow up in more or less stable families, and how this affects their later life labor market and health outcomes. Literature shows that family environment is important for the child's human capital accumulation (e.g. Cunha and Heckman, 2007; Björklund and Salvanes, 2011). Disparities early in life may amplify in adulthood, which can be considered unfair given that a child has no influence over its parents nor its timing of birth. If children

[1]I wish to thank Janet Currie, Anne Gielen and Olivier Marie for useful discussions and support. This paper benefited from comments made by Brian Beach, Leah Boustan, Alex Bryson, Arnaud Chevalier, Gordon Dahl, Emilia Del Bono, David Dorn, Reyn van Ewijk, Ralf Futselaar, Emeric Henry, Hans van Kippersluis, Erik Plug, Frans van Poppel, Maria Rosales-Rueda, Dominik Sachs, Robert Scholte, David Silver, Hannes Schwandt, Tom Vogl, Dinand Webbink, and seminar participants at Erasmus University Rotterdam and Princeton University. All errors and emissions are my own.

are born to parents with different characteristics due to changes in the socioeconomic environment, governments may want to adjust spending in ways to account for both a changed cohort size and composition. My main findings are twofold. First, I find no evidence that better socioeconomic conditions induced by the end of war lead to parental selection. Second, when exploring heterogeneity and focusing on unplanned conceptions that occurred as a response to improved socioeconomic conditions, growing up in a less stable family does not lead to poorer long-term child outcomes.

Fertility responds to economic conditions, and is pro-cyclical as better (worse) labor market conditions lead to higher (lower) fertility (Becker, 1960; Ben-Porath, 1973; Lindo, 2010; Schaller, 2016). An economic downturn affects fertility through an income and substitution effect (Gronau, 1977). A negative income effect leads to a lower demand for children, whereas the time-intensive nature of raising children causes a substitution effect leading to a higher demand for children. The procyclicality of fertility suggests that the income effect is larger than the substitution effect. However, the magnitudes of both effects can differ across families, implying that economic conditions can affect both the size and composition of cohorts. For instance, Perry (2004) shows that for total fertility, the income effect dominates for women with high earnings potential, whereas the reverse is true for women with low earnings potential. In line with this finding, several papers demonstrate that women conceiving during times of high economic uncertainty are negatively selected in terms of education and health (Dehejia and Lleras-Muney, 2004; Del Bono et al., 2012; Currie et al., 2015), which results in lower educational outcomes for their children

(Chevalier and Marie, 2017).[2] Little is known about parental selection in a context of a more severe shock to the socioeconomic environment, which might translate in even larger compositional changes.

This paper exploits such a large change in socioeconomic conditions, and specifically takes advantage of an improvement in the socioeconomic environment induced by the end of the World War II in the Netherlands. The liberation eliminated the poor socioeconomic conditions that prevailed in the last war years as economic recovery was fast and started right after the war (Klemann, 2002). Also living conditions improved with the elimination of war-related uncertainties. The improved socioeconomic circumstances are perhaps best illustrated by the unprecedented rise in fertility that occurred in the Netherlands in 1946 - about nine months after the end of WWII.[3] The Birth Peak represents an unprecedented and temporary increase in the birth rate of approximately 35% from 1945 to 1946. It represents a significant amount of $37,400$ excess births,[4] on $284,000$ actual births. The peak was short-lived as birth rates quickly converged to lower levels, suggesting that changed family formation norms cannot explain the peak (Beets, 2011).[5] The availability of rich Dutch administrative data

[2]Dehejia and Lleras-Muney (2004) and Currie et al. (2015) focus on economic uncertainty in terms of unemployment rates, Del Bono et al. (2012) on job displacements, and Chevalier and Marie (2017) on the transitional period which was characterized by high economic uncertainty after the fall of the Berlin wall.

[3]Other potential explanations for the Birth Peak include a catching up of fertility after the famine in the winter prior to the liberation, migration, and the absence and later abundance of men are addressed in Section 2.2.4.

[4]The difference between the actual and the predicted amount of births. Author's calculations with publicly available data from Statistics Netherlands (statline.cbs.nl). Predictions are calculated by taking the average increase in the number of births from 1941 to 1944 and extrapolating for 1945 and 1946.

[5]It is not uncommon for fertility to respond to changes in the socioeconomic environment induced by war: the Netherlands faced a fertility rise 26 years earlier after the end of the First World War, and fertility surges also arose in other previously occupied countries after WWII (Van Bavel and Reher, 2013). Besides, within the Netherlands dips (peaks) in fertility are observed nine months after worsening (improvements) of war

allows me to study whether being born in a particular family affects labor market outcomes more than 50 years later, and health outcomes more than 60 years later. This combined with the larger increase of the birth rate in the area in which the liberation caused larger changes in the socioeconomic environment, make the Dutch Birth Peak a particularly relevant historical episode to use as a natural experiment.[6]

The regional variation and timing are exploited in a difference-in-difference framework. The results show that children born in regions that faced a larger shock to the socioeconomic environment after the liberation - the Children of the Birth Peak (CoBP) - grow up in smaller families, although no effect is found on the stability of their parents' marriage. In terms of long-term child outcomes, CoBP are not different in terms of employment, labor earnings, and enrollment in social insurance programs. Moreover, no effects are found on mortality before age 65 and age 70. CoBP do have a lower use of prescription drugs in groups relevant to the family environment, which is especially driven by a lower use of prescription drugs for mental health problems. Thus, the cohort is not growing up in more or less stable households, and does not have better or worse labor market outcomes in adulthood. These findings are robust to different definitions of the control and treatment group.

The lack of an effect found may be caused by the fact that two types of behaviors led to pregnancies after the war: (1) couples may have de-

circumstances.

[6]Many countries faced an increase in fertility for some years following the end of the war (e.g. US, Canada, France, Belgium). The birth rate in the Netherlands increased in the year after the liberation, but quickly stabilized to lower levels (Van Bavel and Reher, 2013). Only the birth rate in England and Wales shows a similar behavior (Van Bavel and Reher, 2013). Due to data availability and the presence regional variation I focus on the Netherlands.

layed fertility to better times and (2) unanticipated pregnancies may have occurred in the wave of optimism and celebrations that prevailed after the war. These two types of behaviors may reflect different types of parental selection. Evidence shows that especially women with better socioeconomic characteristics adjust fertility in times with higher economic uncertainty (e.g. Dehejia and Lleras-Muney, 2004; Del Bono et al., 2012; Chevalier and Marie, 2017). Although it is unclear that this mechanism extrapolates to changes in socioeconomic circumstances induced by war, it might be that the children conceived due to delayed fertility have better long-term outcomes. Equivalently, evidence shows that access to family planning measures like the pill and abortion, measures that can prevent unanticipated/unplanned births, are associated with improvements in the labor market outcomes of children born (e.g. Gruber et al., 1999; Ananat et al., 2009; Ananat and Hungerman, 2012; Bailey, 2013; Mølland, 2016). Hence, children born from unanticipated pregnancies are more likely to have negatively selected parents, which may result in worse long-term outcomes. The difference-in-difference strategy would give the *average* differences in outcomes. By using information on the marital states of the parents a distinction is made between these planned and unplanned conceptions. More precisely, when focusing on a sample of first births who are born in-wedlock, a division is made between children who were conceived in-wedlock and children who were conceived out of wedlock (i.e. in shotgun marriages). Given the nature of family formation norms in the Netherlands at the time, the first group is more likely to be a product of delayed fertility, whereas the latter are more likely to be unanticipated.

When estimating the difference-in-difference models for these sub-sam-

ples of planned and unplanned conceptions a different picture emerges. First, when comparing the labor market and health outcomes of planned and unplanned conceptions, children born from unanticipated conceptions have significantly worse labor market and health outcomes. Second, although CoBP who are conceived in-wedlock grow up in smaller families no effects are found on their parents' marital stability. Conversely, for CoBP born in shotgun marriages no effect on family size is found, whereas their parents' marriage lasts on average 5.6% shorter. Hence, within a sample of unanticipated conceptions, those conceived during a period with better socioeconomic circumstances, grow up in less stable households. Third, for the samples of planned and unplanned births no effects are found on various labor market and health outcomes beyond age 50. Summarizing, although CoBP born from unanticipated conceptions grow up in less stable family environments as measured by the length of parental marriage, this does not result in different labor market outcomes and health outcomes in adulthood.

One might be concerned that cohort effects may mask any parental selection effects, as crowding out may negatively affect labor market and health outcomes in adulthood (e.g. Bound and Turner, 2007; Brunello, 2010), which may drive the estimates to zero. First, controls for cohort size on the province level are added to capture potential crowding effects in the classroom as well as on the labor market. Second, I examine the outcomes of the younger siblings of the studied cohorts. These siblings are born to the same mother, but in different times, and thereby this analysis can separate the effect of cohort-specific effects and parental selection. Based on these tests I do not find evidence for cohort effects driving the results found.

Further, throughout my analysis I do not find that family size affects child long-term outcomes. Ultimately, I check how much other explanations for the fertility rise, apart from changed socioeconomic conditions, affect the main findings. More precisely, I focus on the influence of the famine in the winter prior to the liberation, the absence of men during war years, and the presence of Canadian soldiers after the liberation. The tests performed do not find evidence for the contributions of these other explanations, and it appears that the changed socioeconomic conditions are indeed the biggest contributor to the fertility rise.

This paper contributes and improves the existing literature in several ways. First, to the best of my knowledge, this is the first paper that exploits the end of war as a natural experiment to study parental selection and children's outcomes. As such, this paper contributes to the literature on how economic uncertainty affects parental selection (Dehejia and Lleras-Muney, 2004; Del Bono et al., 2012; Currie et al., 2015), and consequently child outcomes (Chevalier and Marie, 2017). The change in socioeconomic circumstances studied in this paper is much larger than business cycle fluctuations studied in other papers. Chevalier and Marie (2017) exploit a shock of a similar magnitude, namely the fall of the Berlin Wall, but study child outcomes that occur much earlier in the life-cycle (i.e. educational outcomes in childhood and adolescence). It appears that the fertility responses that are documented to occur with changes in economic uncertainty do not extrapolate to this particular setting. Suggesting that the influence of the economic environment in the decision making process in the demand for children (as outlined by Gronau, 1977) may work differently in times of war.

Second, this paper adds to the literature on the long-run effects of family environment on child outcomes (e.g. Cunha and Heckman, 2007; Björklund and Salvanes, 2011). It focuses on how the parents' marital stability affects child long-term labor market and health outcomes for over 50 years. It particularly adds to the literature on the long-run effects of being born unanticipated (e.g. Gruber et al., 1999; Goldin and Katz, 2002; Bailey, 2006; Ananat et al., 2009; Mølland, 2016; Myers, 2017). Within the sample of unplanned conceptions I do not find that growing up in a less stable family environment affects long-term child outcomes. This is in contrast with the literature that usually finds long-term effects of growing up in a particular family. The lack of an effect found may suggest that parental marital stability is not related to very long-term child outcomes, but not that there is no effect of family environment on child outcomes. Similarly, the considered outcomes are measured beyond age fifty, suggesting that any adverse effects that may occur earlier in the life-cycle cannot be identified. More research is needed on the effects of changed socioeconomic conditions on different parental characteristics, on outcomes observed earlier in the life-cycle, and on whether these results translate to other settings with comparable large changes in socioeconomic conditions.

This paper proceeds as follows. Section 2 discusses the institutional background of the Birth Peak that is underlying the natural experiment exploited in this paper. The difference-in-difference strategy is discussed in Section 3. Section 4 explains the historical macro-data as well as the administrative micro-data used in this paper. The results on family environment, child labor market and health outcomes are outlined in Section 5. Section 6 explores several mechanisms that could be driving the results

found and a conclusion is provided in Section 7.

2.2 The Dutch Birth Peak of 1946

2.2.1 The Netherlands in WWII

Germany invaded the Netherlands in May 1940 after the start of World War II in September 1939. Clearly the Dutch were not pleased with the German occupation, but they had no other choice than to adjust to the new situation. The collaborative nature of the Dutch led to extra trade with the Germans, and as a result the Dutch economy was booming in the first 1.5 years (1940/1941) of the occupation. Unemployment disappeared, there was little poverty, farmers and store-owners were doing well, and companies made large profits. In these first years the Dutch barely resisted against the German occupation. This changed when the Dutch economy collapsed by the end of 1941. The years thereafter were characterized by impoverishment, 'redundant' companies were closed by the Germans, and employees were exported to Germany. Klemann (2002) argues that the impoverishment that prevailed after 1941 coincided with an increase in the resistance against the German occupation. The resistance increased even more with the turning point of the war[7] after which people started to believe Germany would eventually loose the war.

The allied forces initiated the attempt to liberate Western Europe in the spring of 1944. The south (Figure 2.1) of the Netherlands was liberated by the Allied forces in September 1944. The attempt to liberate the northern

[7]The turning point of the war came with the losses of Nazi Germany in Stalingrad (August 1942 to February 1943), Kursk (July-August 1943), and North-Africa (June 1940 to May 1943).

provinces failed, and the Dutch government called out a railroad strike to support the Allied forces in their attempt to free the nation. The occupying forces reacted by enforcing an embargo on food transport to the densely populated west, which caused a famine in its large cities (denoted by black dots). A new liberation attempt was initiated by the Allied forces in the spring of 1945, and the Netherlands was liberated on the 5th of May 1945 with the surrender of Germany.

The liberation ended five years of war, took away uncertainties associated with war times and improved socioeconomic conditions. The changed socioeconomic environment is best reflected by the unprecedented rise in the number of births about nine months after the liberation.

2.2.2 The Birth Peak

The yearly crude birth rate (number of births per 1000 individuals) in the Netherlands is documented in Figure 2.2. The Netherlands is a small country and is very much depending on its international connections. Protectionist policies played a big role after the First World War, and the Dutch economy suffered. This explains, together with the onset of the Great Depression in the 1930s, why the birth rate was decreasing in the 1920s and early 1930s. The Dutch government abolished the Gold standard in 1936, which should have started economic recovery. However, as major trade allies were forming alliances with other countries, it was hard to catch up economically, and the birth rate remained low up until the first war-years. The start of the war in 1939 and the invasion of the Netherlands in 1940 caused an inflow of trade from Germany, leading to a booming economy. This economic expansion and low unemployment increased the birth rate

right from the early war years. The birth rate kept increasing during the war, despite the economic downturn that started around 1942. Pro-family policies, i.e. tax benefits and child subsidies, were implemented by the government during the war (Klemann, 2002), which may explain the upward trend in the birth rate.

What stands out in Figure 2.2 is the unprecedented and temporary increase in the birth rate following the end of WWII (vertical lines) up to 30 births per 1000 inhabitants.[8] The monthly number of births started to increase rapidly from March 1946 and the largest number of children were born in May 1946. Given that the Netherlands was officially liberated in May 1945, the birth rate started to increase 10 months after the liberation.[9] The yearly birth rate increased by 35% from 1945 to 1946, after which it quickly returned to lower levels.[10] The peak represents an unprecedented rise in the number of children born as never before this many children were born in a year.[11,12] The quick reversal to pre-liberation levels suggests that it cannot be explained by permanent changes in family formation attitudes (Beets, 2011), and it is rather a response to the liberation. Although the liberation might have been expected with the weakening of the German occupants, the timing of the fertility response about nine months after the

[8]A crude birth rate of 30 is relatively high when comparing it to other countries in the same year. A similar high crude birth rate can be observed for Poland after WWII. Comparable but somewhat lower birth rates can be observed for Canada, Finland and Iceland in that same year (Van Bavel and Reher, 2013).

[9]This is not surprising as it took until the beginning of June before all Dutch municipalities were actually free of German occupants.

[10]This number ignores the dip in birth rates in 1945. The birth rate increased by approximately 29% from 1944 to 1946.

[11]Before 1946, the largest number of children were born in 1920, two years after the end of the first World War (vertical lines).

[12]The peak cannot be explained by a late registration of children born during the war, as this would have resulted in a rise in 'reported' births right after the liberation in May 1945.

liberation suggests that it was not anticipated.

The south of the Netherlands was liberated in September 1944, whereas the northern provinces were liberated in May 1945. Implying that if the Birth Peak is the result of the liberation and changed socioeconomic conditions there should be geographical differences in the magnitude of the Birth Peak. Apart from the differential timing of the liberation in different parts of the Netherlands it is important to take into account that dense-populated cities in the west suffered from a famine in the winter prior to the liberation. This allows for the identification of three different areas which are depicted in Figure 2.1. First, the south that was liberated in 1944 and was not affected by the famine, (2) the north that was liberated in 1945, and (3) large cities in the west that suffered from the famine and that were liberated in 1945.

Figure 2.3 shows the fertility trends for the north (in- and excluding the large cities in the west) and the south. What stands out is that other than differences in the levels, the trends look remarkably similar across regions before 1945.[13,14] The magnitude of the fertility response is largest in the north, and practically absent in the south. A simulation exercise is done to illustrate the differences in magnitudes. When extrapolating the growth in the number of births from 1941 to 1944 for 1945 and 1946, the observed number of births is higher than the predicted number of births by $17,092$ in the north (excluding famine-affected areas), and $2,815$ in the

[13]Historically the south is known for its Roman-Catholic denomination, whereas the Protestant and Calvinist faith are the main religions in the north. Fertility is higher among Roman-Catholics as opposed to other religions, and fertility is higher in rural versus urban areas (Engelen, 2005). This explains the different levels of fertility among the Catholic and rural south, the rural northeast, and the urban northwest.

[14]The areas are also different in population size. About 52.1% of the Dutch population is based in the north (when excluding the large cities in the west), 25.6% is located in these large cities in the west, whereas the remaining 22.3% is based in the south in 1946.

south. Implying that the end of the war contributes to the birth rate by 3.5 extra births per 1000 inhabitants in the north, and 1.3 for the south. The peak in the birth rate is also visible when focusing on the number of births (Figure 2.A1).

The dip in birth rates in 1945 can be explained by the onset of the famine. Children born in famine-affected areas are excluded from the analysis to prevent that famine-exposure is driving fertility and parental selection after the war.[15] Leaving out those children decreases the magnitude of the dip from -9.2% to -4.4%.[16,17] Figure 2.4 shows the yearly percentage change in the number of births for north and south. The change in the number of births followed a similar pattern in north and south prior to 1945. What stands out is that the growth in the number of births is particularly large in 1946 for the north (+33.0%) whereas the change is much smaller in the south (+8.5%).

The explanation for the slight impact for the south is less straightforward. The south was liberated in 1944, and if anything a similar fertility response would be expected to occur in 1945 given a pregnancy duration of nine months. First, more than half of the country was still occupied by Nazi Germany after the liberation of the south, which may lower the responsiveness of fertility to the liberation. Second, the south was characterized by a large share of Catholics who had traditional family formation norms, which might translate in less responsive fertility.[18] Third, family

[15]That is, all children born in large cities in the west (i.e. Amsterdam, Delft, Haarlem, Leiden, Rotterdam, The Hague, and Utrecht), which is about 25.6% of all births in 1946.

[16]An extensive discussion on the potential effects of the famine and how I deal with this is provided in Section 2.2.4.

[17]The decrease in the number of births from 1944 to 1945 for the north including, and excluding the large cities in the west.

[18]The Dutch Catholics had developed a certain rigor after having had a minority position for years (Van Poppel, 1985). Out-wedlock births were highly undesirable at

size was on average larger in the south as opposed to the north, which may also make fertility less responsive to changes in war conditions. Hence, the regional variation between the different regions is likely caused by a combination of family formation norms and the fact that more than half of the country was still occupied. The latter implying that the change in socioeconomic conditions was probably smaller in the south at the time of its liberation.

2.2.3 Delaying and/or celebrating?

The liberation also affected other family structure decisions, such as the marriage rate (Figure 2.5).[19] Marriage and fertility decisions were strongly intertwined half-way 20th century Netherlands. Marriage took place whenever the couple gathered sufficient material and financial resources. Children were generally born in-wedlock,[20] i.e. within a marriage, and premarital conceptions were followed by shotgun marriages such that the child was born in-wedlock. During WWII stark spikes in the number of marriages are visible in 1939, 1942 and 1946,[21] whereas a similar trend deviation for fertility is only observed in 1946.[22] Suggesting two types of behaviors that led to pregnancies in 1946. First, individuals who got married in 1939 and 1942 (and the years in between) delayed conceptions up until after the war. Second, the peak in marriages and fertility in 1946 suggest that premarital

the time, but the social stigma associated with a premarital conception followed by a marriage was completely different. For Catholics both types of premarital conceptions were unacceptable.

[19]The same Figure for the number of marriages is shown in Appendix Figure 2.A2.

[20]Out-wedlock births comprised only 2.5% of all births in 1946. The percentage of out-wedlock births is lower in surrounding years (Statistics Netherlands).

[21]Regional data is missing for 1940, but on the country level a decrease in the number of marriages is observed in 1940.

[22]The effect of an absence of men during war times on fertility, and hence perhaps also on the marriage rate, is discussed in Section 2.2.4.

conceptions, occurring in the wave of post-war optimism, led to shotgun marriages.[23] Given family formation norms in the Netherlands, it might well be that shotgun marriages indicate that the conception may have been unanticipated. The fertility rise appeared for marriages of both shorter and longer duration which suggests that both mechanisms are at play (Van den Brink, 1950).

Figure 2.6 shows the trend in maternal age at first birth by region using the micro-data. Note that the line for the north excludes famine-affected areas, which are plotted separately. The trends look very similar before and after 1946 for the north and south. The trends in maternal age at first birth can be explained by economic uncertainty, where higher (lower) economic uncertainty leads to lower (higher) fertility (Becker, 1960; Ben-Porath, 1973; Lindo, 2010; Schaller, 2016). Maternal age at first birth was increasing up until the end of the war, which can be explained by the slow recovery from the Great Depression. Bad (socio-) economic circumstances complicate the process of marriage and childbearing as it is hard to gather sufficient financial and material resources when unemployed. The trends in maternal age at first birth reversed after the first post-war years. The economic circumstances were good after the war so there were no reasons to delay marriage. Female labor market participation increased, making it easier to obtain sufficient resources for marriage (Van Poppel and Willekens, 1982). This led to a decline in age at first birth of about two years in the after-war years (CBS, 2012).

[23]These shotgun marriages were quite common in the Netherlands, respectively 11.6% and 13.4% of marriages in the marriage cohorts of 1945 and 1946 had a birth within 6 months after marriage (Van den Brink, 1950). Figure 2.A3 shows the higher incidence of out-wedlock conceptions after the liberation, which is driven the higher number of shotgun marriages.

What stands out in Figure 2.6 are the jumps in maternal age at first birth in 1946. Maternal age at first birth can be interpreted as an indicator for fertility delay as well as unanticipated pregnancies. Parents may have delayed their first birth until better times. At the same time, unanticipated pregnancies are more likely to occur for single unmarried women, and hence in first birth. A jump in maternal age at first birth is expected for the 1946 birth cohort if women delay their first birth due to the war. This increase may be dampened by the occurrence of unanticipated pregnancies, which generally occur at lower ages.[24] The increase is larger in the south, which may be explained by smaller incidence of unanticipated pregnancies.[25] Figure 2.7 shows the number of months between marriage and first birth, which is another indicator for fertility delay. This difference will be larger after the liberation if married couples delayed their first birth. The difference is increasing up until 1945 for the south, which is consistent with their liberation in 1944. For the north, the difference in increasing up until 1946, which is also consistent with the end of the war in 1945. Suggesting that married couples did delay their first birth until better times.

Figure 2.8 shows the trends in maternal age at second birth by region. A jump in maternal age at second birth is visible in the north and south in the year after their respective liberations (i.e. 1945 for the south and 1946 for the north). Hence for second births, in which parents were probably already married, there is a fertility response in the year right after the liberation. Suggesting that parents delayed second births until better times, and given that they were already married they could start right away. This might

[24]Maternal age at first birth in 1946 is 4.7 years lower for shotgun marriages as opposed to in-wedlock conceptions (22.8 versus 27.5) in the marital sample ($P = 0.000$).

[25]Also women in the south were on average older when having their first child, and thereby pressure to have a child after the war may have been higher.

also explain the absence of a fertility response for first births in 1945 for the south. As mentioned before, the south had very conventional family formation norms at the time due to its mainly Catholic denomination. Suggesting that fertility is less responsive because children were conceived within marriage.

Together the evidence above illustrates that (a) the trends in fertility in north and south were very similar before 1945, (b) the fertility rise in the north was sudden, temporary, and of unprecedented magnitude, and (c) the trends in the marriage rate, maternal age at birth, and the timing from marriage to first birth suggest that both delayed fertility and unanticipated conceptions played a role during the Birth Peak.

2.2.4 Explaining the Birth Peak

Four potential explanations for the Birth Peak are: (1) the catching up after the Hunger Winter, (2) migration, (3) the relative absence and later abundance of men, and (4) improved socioeconomic conditions. The changed socioeconomic conditions after the liberation seem to be the greatest contributor to the fertility rise and are the focus of this paper. The relative importance of each explanation and strategies to deal with potential problems are discussed below.

'Catching up' after the famine

Large cities in the west of the Netherlands suffered from the Hunger Winter in the winter prior to the liberation. A lack of food affects the ability to conceive. To illustrate, during the famine 16% of women had irregular menses (Elias et al., 2007) and half of Dutch women were not menstruating

during the famine (Roseboom, 2010). Additionally, women who lost a child during the famine may be more likely to bear another child after the famine (e.g. Nobles et al., 2015). Hence, the fertility response after the liberation could be perceived as a 'catching up' of fertility after the famine. However, it cannot be the primary explanation for the peak. Figure 2.3 shows a fertility rise even in absence of the large cities in the west.

The famine could be problematic for the set-up of control and treatment groups. If women in famine-affected areas are more likely to have experienced conception difficulties during the famine, or if maternal health is affected by the famine, this may impact parental selection and the outcomes of their children. Therefore, children born in exposed areas, i.e. those born in the large cities in the west, are left out of the treatment group. Leaving these children out will likely capture most of the potential effect of the famine on subsequent fertility. Food rations show that the famine only played a role in the west (Table 2.A1). Within the west, only urban areas suffered as people could produce food themselves on the countryside. To illustrate, when people in The Hague and Leiden were starving, villages nearby were much better off (Stein et al., 1975). Similarly, mortality during the famine was specifically high for males in the urban west, whereas mortality in the rural west and the rest of the country was very comparable (Ekamper et al., 2017). Hence, dropping children born in large cities in the northwest will probably eliminate the impact of the famine.[26] However, Figure 2.3 shows a small dip in fertility even when leaving out these cities. I execute robustness checks in which I leave out all children

[26]In their study on the in-utero impact of the famine Scholte et al. (2015) and Stein et al. (1975) children are exposed to the famine if they are born in these large northwestern cities.

conceived during the famine which does not change the results.

Migration

Migration could be responsible for the Birth Peak in the north especially if pregnant women moved to the north. Freely moving across the country was difficult during the German occupation, but also right after the liberation. The infrastructure suffered as roads, railroads, and bridges were destroyed during the war. This is particularly important as the northern and southern provinces are separated by rivers, and bridges were key means of transportation from the south to the north. Figure 2.9 shows net migration[27] per 1000 inhabitants by region.[28] If migration was the main driver of the Birth Peak in 1946 one would expect positive net migration in the north (excluding the big cities in the west), which is not the case when looking at Figure 2.9. Hence, it is unlikely that migration is driving the fertility rise. Moreover, Figure 2.3 controls for migration by taking into account the average[29] yearly number of inhabitants in the region in the denominator, and the presence of a peak suggests that migration cannot be the main driver. Migration might especially pose problems for famine-affected areas, which is shown in Figure 2.9 as the line including these areas shows a net inflow of people after the war, as this is where circumstances were worst prior to the end of the war. For other reasons discussed above children born in famine-affected cities are left out of the analysis, which would also address this potential issue.

[27]The number of immigrants minus the number of emigrants.

[28]A figure with net migration (not per 1000 inhabitants) is shown in Appendix Figure 2.A4.

[29]The average of the number of inhabitants at the beginning of the year and the number of inhabitants by the end of the year

From absence to an abundance of men

The absence of men during war-years can lead to lower fertility as conceiving is hard when men are away, and similarly their return might cause an increase in conceptions.[30] There was basically no mobilization of the army in the Netherlands, as the Dutch army fought for four days only.[31] However, forced labor did generate a big loss of young and fit manpower (Stein et al., 1975). About 531,000 men were recruited for forced labor in Germany during the war years, and approximately 8,500 died (Krimp and Kemperman, 2015).[32] Considering that the population of the Netherlands aged between 20 and 45 in 1945 was about 3,462,000, and by assuming a sex ratio of 0.5, this would give an approximation of 1,731,000 men in childbearing ages. Hence, about 31% of men in 'childbearing ages', i.e. aged between 20 and 45, were in forced labor at one point during the war.

Even though forced labor entailed a considerable temporary loss of men in childbearing ages, there is no indication for a non-proportional withdrawal of men across the country prior to the winter of 1944. Especially men working in companies that were closed by the Germans, and those in specific age-categories were called up for forced labor (Sijes, 1990). The fall of 1944 was characterized by raids that took away a significant amount of men in especially the cities of Rotterdam and The Hague (to illustrate, in Rotterdam about 50,000 men were taken away in November 1944). Child-

[30]The relatively abundance of men, i.e. the sex ratio, also interacts with fertility. Higher sex ratios lead to better bargaining positions for males (Angrist, 2002; Abramitzky et al., 2011; Bethmann and Kvasnicka, 2013; Porter, 2016).

[31]They fought from 10 to 14 May 1940 and 2,200 to 2,300 soldiers were killed in action.

[32]Other war casualties in the Netherlands include about 100,000 Jews, 16,000-25,000 victims from the Hunger Winter, 50,000 deaths due to lower public health, 30,000 civilian casualties of war, and 23,000 civilian casualties of the liberation (Krimp and Kemperman, 2015).

ren born in these two cities are left out of the analysis to prevent that the absence of men is driving the fertility response after the liberation. Overall, it is unlikely that the absence of men is driving the fertility rise in 1946 as even though many men were away during the war, fertility was still increasing during the war and responding to changes in war circumstances (Figure 2.11).

Another explanation that has been put forward for the post-WWII baby-boom in the US is increased female labor force participation during the war due to the absence of men (Doepke et al., 2015). Women who were sufficiently old to work entered the labor market during the war and gained labor market experience. After the war, younger women with less labor market experience were crowded out of the labor market, especially with the return of the men. Fertility was an outside option for these younger women, and the birth rate increased for them. For the Netherlands this explanation seems very unlikely. First, women did not disproportionately enter the labor force. And second, the Dutch fertility response is driven by older women, which is in contradiction with this hypothesis (see Figure 2.A5).

After the war there was not only an inflow of Dutch men back into the country, there was also an inflow of Allied forces soldiers into the country. The Netherlands hence moved from an absence to a relative abundance of men. The festivities with the Canadian Allied forces led to the birth of approximately 7,000 'libertarian babies' (estimate from Okkema, 2012).[33] The majority of these babies were conceived after the liberation of the south, which explains the higher number of out-wedlock births in 1945 as

[33]Soldiers from the US, the UK, and Poland also contributed to the liberation. However, the focus of Dutch media solely on the involvement of Canadian soldiers with Dutch girls, suggests that the others played only a minor role.

opposed to 1946 (see Figure 2.A3). A conservative estimate of the amount of babies born to the Allied forces in 1946 would be half of the total amount born to the soldiers, 3,500. This is 1.2% of the total amount of births in 1946, and 9.4% of the number of excess births (i.e. 37,400). Hence, the presence of the Canadian soldiers increased births but cannot be the primary explanation of the Birth Peak. The mechanisms section explores the sensitivity of the results to the presence of the Canadian liberators by exploiting two strategies. First, the analysis is run with children conceived after the largest departures of Canadian soldiers. Second, children born in municipalities with closer proximity to the Canadian soldiers are left out of the analysis.

Improved socioeconomic conditions

Fertility is affected by economic uncertainty as better (worse) economic conditions lead to higher (lower) fertility (Lindo, 2010; Schaller, 2016; Chevalier and Marie, 2017). The Dutch economy was not doing well after the Great Depression in the 1930s. The start of the war initiated trade with the Germans which improved economic conditions. Figure 2.10 shows the trends in GDP from total production and industrial employment. The Dutch economy was doing well in the first 1.5 years of the occupation but eventually collapsed by the end of 1941, which is shown by declines in industrial employment and GDP between 1942 and 1945. Post-war economic recovery was fast and kicked in right after the war.

Bad economic conditions caused the birth rate to remain low after WWI and recovery started with the improved economic conditions in the first years of the occupation. The birth rate remained high during the

war, which may be explained by the pro-family polices implemented at the time. The end of the war came with improved economic conditions, but the effect on society may have been even larger. The liberation did not only improve economic conditions, but entailed a change in socioeconomic and living environment as times of war were over. The initiator of WWII and the leader of Nazi Germany had committed suicide and the war was over. Even though combat action did not affect all civilians, the war was associated with poor nutrition, and high stress levels due to the risks of persecutions, bombings or combats (Lindeboom and Van Ewijk, 2015). These uncertainties were eliminated with the end of the war. The significance of the change in socioeconomic conditions that came with the liberation is illustrated by peaks in fertility that occurred earlier for countries that were not affected by the war. This recovery took place after the end of the war for involved countries, suggesting that the liberation was a sufficiently large shock to the socioeconomic environment (Van Bavel and Reher, 2013). This is confirmed by the timing of the fertility rise about nine months after the liberation.

The responsiveness of fertility to socioeconomic conditions is also demonstrated by the monthly crude birth rate (Figure 2.11). There are drops in fertility nine months after the start of the war in September 1939, nine months after the German invasion, nine months after the February strike of 1941, and nine months after the Hunger Winter of 1945/55.[34] Birth peaks were observed in March 1945 (nine months after the invasion of the Allied forces in Normandy), and about nine months after the liberation. Both

[34]The fertility dip after the famine in 1945 seems to be balancing the fertility rise in 1946. However, this picture does not take into account regional variations in the fertility rate, as these are unfortunately unavailable.

birth peaks can be explained by optimism surrounding the expectation of favorable war circumstances (CBS, 2012). The general responsiveness of fertility to changes in war circumstances together with the timing about nine months after the liberation suggest that changed conditions were driving the 1946 Birth Peak.

2.2.5 What about the 'marginal child'?

The 'marginal child', the child that would not have been born in the absence of post-war optimism, could be different from the average child through two mechanisms. Section 2.2.2 established that the fertility rise is driven by both delayed fertility and unanticipated conceptions. Parents who delayed fertility are likely different from those who faced an unanticipated conception, and so may be their children's outcomes. First it is important to stress that parental selection after the war is not driven by the availability of medical care. The medical sector expanded during war years.[35] Equivalently, the number of perinatal deaths and stillbirths is not higher for the 1946 cohort, whereas child mortality is somewhat elevated (Figure 2.A6). This suggests that health care conditions were sufficient, and that parental selection is driven by choices instead of available care. If anything the higher prevalence of child mortality would hint at negative parental selection for this cohort.

First, the bad socioeconomic circumstances that prevailed during the last war years, together with the expectation that the Germans would eventually be defeated after the turning point of the war could cause parents to

[35]The number of doctors, dentists and midwifes increased from 1938 to 1942 by respectively 5%, 28% and 10%. In 1948, 22% more people were employed in the medical sector as opposed to 1938.

delay fertility until after the war. Previous literature on parental selection and economic uncertainty shows that women with better socioeconomic characteristics are more likely to respond to economic uncertainty by adjusting fertility (Dehejia and Lleras-Muney, 2004; Del Bono et al., 2012; Currie et al., 2015; Chevalier and Marie, 2017). This would imply positive parental selection for parents who delayed fertility during the war, and consequently better long-term outcomes for the 'marginal child'. However, during a war it is unclear who is delaying fertility. It is hard to survive with the available means, and people had to use their assets to survive. Richer individuals have greater resources than poorer (in terms of money and goods) to exchange for necessary goods, which might lead to a larger impact of the war on individuals from lower socioeconomic strata during the war. However, the low social classes (i.e. laborers) were not necessarily worse off as they often started trading on the black market (Klemann, 2002). Hence, it is not a priori clear who is delaying in times of war.

Apart from the ambiguity of the interaction of war circumstances and parental selection, fertility delay is not by definition good for the child. Older mothers biologically have a higher probability on adverse pregnancy outcomes (e.g. Abdalla et al., 1993; Gianaroli et al., 1999; Pellicer et al., 1995; Ananth et al., 1996; Stein and Susser, 2000), which may also impact the child's later life outcomes. These adverse health outcomes become apparent beyond certain age, likely around age 33-35 (e.g Royer, 2004; Miller, 2011; Bratti and Cavalli, 2014). This biological channel will likely only play a minor role in this paper, as the margin of later motherhood is centered around age 27 (see Figure 2.6 and 2.8). Additionally to take into account any biological concerns, I control for maternal age at birth in the empirical

strategy. Hence, the biological channel will likely not affect the results. Similarly delaying fertility might be good as it provides women with the opportunity to invest in their human capital before childbirth which improves labor market outcomes (Miller, 2011; Bratti and Cavalli, 2014), and can be beneficial for the child (Miller, 2009). However, this channel mainly works through increased parental resources and home stability induced by delayed fertility (Fergusson and Woodward, 1999). This channel may play a smaller role in the studied time period as human capital accumulation may be complicated during war times, and either way female labor force attachment was very low in the 1940s/1950s.

Second, immediate post-war optimism could lead to unanticipated conceptions. Women might get pregnant in the wave of optimism that prevailed right after the war without well considering the consequences. This is especially relevant there was no access to oral contraceptives at the time, and induced abortion was illegal. Despite the availability of other (less effective) contraceptives (i.e. rubber condoms and periodic abstinence) unintended conceptions were prone to occur, and prone to end up as unanticipated/unplanned births. Earlier literature shows that children whose mothers got (improved) access to abortion, a measure that can prevent unanticipated/unplanned births, have better living circumstances and adult labor market outcomes (Gruber et al., 1999; Ananat et al., 2009; Mølland, 2016). Similarly, improved access to the pill provided women the ability to plan pregnancies and is associated with a lower number of unwanted births (e.g Goldin and Katz, 2002; Bailey, 2006; Myers, 2017). The diffusion of the pill led to positive parental selection in the longer-run (Ananat and Hungerman, 2012), which extends towards better educational and la-

bor market outcomes for their children (Bailey, 2013). The micro-data shows that marriages of parents who conceived in-wedlock lasted on average 7 months longer than marriages of parents who conceived out-wedlock ($P = 0.000$).[36] Suggesting that children born in the latter marriages may grow up in a less stable household environment. Unanticipated pregnancies are hypothesized to be associated with negative selection into parenthood, and the 'marginal child' would have worse characteristics on average.

Last but not least, when studying child adult outcomes in this particular setting it is important to take into account that the studied cohort may not only be different in composition, but is for sure different in size. Evidence shows that cohort size may negatively affect a cohort's educational and labor market outcomes (e.g. Bound and Turner, 2007; Brunello, 2010). For the Birth Peak cohort, the entrance of large groups of pupils into primary schools led to large classes (48 to 50 students were not uncommon in primary school). The cohort entered the labor market in the 1960s. Unemployment was low because the cohort's labor market entrance coincided with a large demand for labor, and staying longer in school became a new option (Schuyt and Taverne, 2004; CBS, 2012). The large size of the cohort is hypothesized to negatively affect adult outcomes. Section 2.6.1 discusses two methods to separate parental selection effects from cohort size effects, and based on these tests I find no evidence for cohort effects.

[36]Based on in-wedlock births in 1946 who are observed in the marital data.

2.3 Empirical strategy

2.3.1 Parental selection during and after WWII

This paper examines how improved socioeconomic circumstances at conception affect the size and composition of a cohort. It exploits regional differences in the timing and magnitude of changed socioeconomic induced by the end of WWII in the Netherlands as a natural experiment. The geographic variation allows for an application of a difference-in-difference strategy to examine parental selection and consequently the outcomes of their children. Studies investigating historical events often compare children born before and after the event to children exposed to the event (Lumey et al., 2011). It is more complicated in this setting as children born before 1946 are exposed to the war. Parental selection will likely be different during the war, directly after the war, and when the situation is back to normal. A hypothetical representation of parental selection responses to war circumstances and their expected effects on child outcomes are depicted in Figure 2.12. The timing in the figure corresponds to the month and year of birth of the child, and for convenience assumes a pregnancy duration of nine months.

The ratio of child outcomes (k) increases with positive parental selection in the north, i.e. if child outcomes improve in north versus south, and the other way around. The ratio is constant in segment A, as both areas are exposed to war and experience similar parental selection. Part B represents the situation after the liberation of the south in September 1944. Ratio k declines when assuming positive parental selection for the south in this

period.[37] The post-war conceptions are represented by segment C. When assuming positive parental selection for the Birth Peak cohort, ratio k would increase with better outcomes for children in the north. Ratio k returns to its steady-state in segment D.

The main test resulting from Figure 2.12 is comparing segment A and B to segment C. As this paper studies how an improvement of socioeconomic conditions shapes parental selection and children's outcomes, the relevant test is to compare children on the margin from worse to better circumstances, that is before and after the liberation. The identification strategy accounts for other war events that can affect parental selection after the liberation. Segment C and D are compared in a robustness check. Both cohorts born after the war are exposed to better socioeconomic conditions, for the latter group no Birth Peak is observed. The first test captures the influence of changed socioeconomic conditions at conception on the size and composition of a cohorts.[38]

2.3.2 Identification strategy

In my analysis I exploit the end of war as a natural experiment that caused an exogenous shock in socioeconomic conditions that induced a shock in the birth rate. The temporary nature of the shock allows for an identification of the cohorts born before and after the liberation. Furthermore, a larger effect on the birth rate is observed in the area in which the liberation caused larger changes in the socioeconomic environment. Hereby the liberation had a large effect on the birth rate in the north, whereas

[37] This figure ignores parental selection caused by the famine as children born in famine-exposed cities are left out of the analysis.

[38] At the same time, selection into the sample after the Birth Peak (e.g. those born in 1947/1947) may be affected by fertility decisions of parents during the Birth Peak.

there was barely an effect on the birth rate in the south. The south serves as a natural control group, and controls for a shared macro-environment that may affect the long-run outcomes of the studied cohorts. I employ a difference-in-difference strategy in which I compare cohorts born before and after the liberation, where the latter are exposed to a better socioeconomic environment, in the north and south. The birth rate started to increase from March onward and remained fairly high up until September 1946 as is shown in Figure 2.12. Therefore, the Birth Peak cohort is defined as born between March and September 1946.[39] The control groups contains children born between January 1944 and February 1946. As of the limited influence of the Birth Peak in the south, I can credibly estimate the effect of the fertility shock while controlling for common cohort-specific effects that may affect long-term child outcomes.

Equation 2.1 is estimated for different outcome variables y_{irt}, where subscripts refer to individual i born in region r, and month/year t. The labor market outcomes of interest are labor force participation, earnings and enrollment in any disability insurance scheme in 1999. The health outcomes of interest are indicators for mortality before age 65 and 70, and an indicator for whether the individual had any prescription drugs for diseases related to the individual's lifestyle (i.e. mental health, cardiovascular, respiratory, and diabetes) in 2006.[40]. Equation 2.1 contains an indicator for being born in the Birth Peak cohort (March-September 1946) (BP_{it}), an indicator for being born in the North of the Netherlands ($North_{ir}$), and their interaction indicating that the child is born in the north during the

[39]Robustness checks with respect to the choice of this definition are provided in Section 2.5.4.

[40]More information on the outcome variables is available in Section 2.4.2.

Birth Peak and hence is considered a Child of the Birth Peak, henceforth CoBP ($CoBP_{irt}$).[41] An indicator for birth in the time period between the liberation of south and north is added ($LibSouth_t$). Linear and quadratic region-specific age trends are added to take into account region-specific age profiles in outcomes. For example, individuals in different regions may sort into different occupations, i.e. the agricultural south and more urban north. Vector $\mathbf{X}_i$ contains individual-specific controls, standard errors (ϵ_{irt}) are clustered by month of birth and region.[42]

$$\begin{aligned} y_{iprt} = \gamma_0 + \gamma_1 BP_{it} + \gamma_2 North_{ir} + \gamma_3 (CoBP_{irt}) \\ + \gamma_4 LibSouth_t + f(MoB, YoB)_{ir} + \mathbf{X}_i \delta + \epsilon_{irt} \end{aligned} \tag{2.1}$$

First, it is important to note that the studied cohorts were all subject to the same educational system, as there were no educational reforms up until 1968 (Dodde, 1983). Likewise, fertility after the war is not driven by the availability of medical care. However, a potential concern with the aforementioned identification strategy is that it is hard to eliminate the impact of other war events on fertility and parental selection. Although children exposed to the Hunger Winter are born from February 1945 to December 1945 (Scholte et al., 2015), i.e. before the Birth Peak, there might be responses to famine circumstances that might affect fertility and selection after the liberation. To exclude any potential confounding by the Hunger Winter all children born in famine-exposed cities (more than $40,000$ inhabitants in 1944)[43] are excluded from the treatment group. Second, to

[41] Hence this indicator represents $BP_{it} * North_{ir}$.

[42] Equation 2.1 is estimated by OLS for the ease of interpretation, but results are robust to the use of binary probit models. Results are available on request.

[43] This is treatment definition of famine-exposure used in Scholte et al. (2015).

exclude that the earlier liberation of the south is differentially affecting fertility across areas, an indicator for births within this time-interval is added to the specification. Section 2.5.4 shows that the results are robust to excluding all children who are conceived during the famine, and during the liberation of the south.

Equation 2.1 cannot distinguish between delayed fertility and unanticipated conceptions, and rather gives the *average* effect of being born in this Birth Peak cohort. By using information on the marital status of the parents, a distinction is made between children who are conceived in-wedlock and those who are conceived out of wedlock. More precisely, to take into account potential premature births, children born within seven months after marriage are defined as in-wedlock conceptions. The difference-in-difference model is estimated separately for first births who are born in wedlock, where a distinction is made between in- and out-wedlock conceptions. The first are more likely to be a product of delayed fertility, whereas the latter are more likely to reflect unanticipated conceptions.

Another worry is that it is hard to distinguish between the effect of cohort size and cohort composition, especially when studying the child's outcomes in adulthood. Two tests are done to explore the effect of cohort size on the results. Firstly, a control variable for the size of the child's birth cohort at the province level is added to the specification to account the potential effects of cohort size. The province level size of the birth cohort is a reasonable proxy for capturing cohort size considering the low residential mobility of the Dutch. Secondly, a stronger test is performed in which I examine the outcomes of younger siblings. Given that these younger siblings are born to the same mother, but in a different time, parental

selection effects can be separated from cohort effects (see Section 2.6.1).

The key assumption of a difference-in-difference strategy is that the trends in outcomes, for both treatment and control, would be the same in absence of treatment (Angrist and Pischke, 2008). This paper is about parental selection, and it is of primary interest to check that parental characteristics in north and south exhibit similar pre-trends. Figure 2.6 shows that the trends in maternal age at first birth are very similar in north and south prior to 1946, and Figure 2.8 shows a similar picture for age at second birth prior to 1945. Figures 2.A9 and 2.A10 show pre-trends at the child level for labor market and health outcomes. Pre-trends look very similar for the outcomes considered prior the last war years. I also test the common trend assumption for parent and child outcomes more formally. Following Autor (2003) I estimate Equation 2.2. The outcome of interest (y) of individual i born in year t and region r is regressed on a set of region fixed effects (λ_r) and year fixed effects (δ_t), standard errors are clustered by birth month/year and region. Indicators D_{it} represent interactions between the treatment variable, i.e. born in the north, with birth year (ranging from 1941 ($j = -5$) to 1950 ($j = 5$) where 1940 is the reference category). The common trend assumption holds if parameter estimates for earlier years are not significantly different from zero. The results are reported in Table 2.A3 and 2.A4. I find no evidence for differences in pre-trends on the parental level and on child level in the years prior to the last war years.

$$y_{irt} = \lambda_r + \delta_t + \sum_{j=-5}^{+5} \beta_j D_{irt} + \epsilon_{irt} \tag{2.2}$$

2.4 Data

2.4.1 Macro-data: historical regional demographics

Province-level demographics and monthly fertility data dating back to before WWII are retrieved from the historical collection of Statistics Netherlands (available at *historisch.cbs.nl*). This online portal contains scans of the original publications from the archives, and I digitized the data in these records manually. Information on the monthly crude birth rate is available in the Statistics of Population Development records (*Statistiek van de Loop van de Bevolking, 1938, p45*). Municipality level data on the number of live births, population at the beginning and end of the year, migration, and number of marriages is available in the records that document population development by municipality (*Loop van de Bevolking per Gemeente, 1930-1950*).

With the above mentioned data I calculate the crude birth rate in year t as the number of live births per 1000 inhabitants.[44] For migration I calculate net migration (incoming minus outgoing) per 1000 inhabitants in year t.[45]. The average population is the average of the population on January 1 and the population on December 31.[46]

[44] $CrudeBirthRate_t = (\frac{LiveBirths_t}{AveragePop_t}) * 1000$

[45] $Migration_t = (\frac{NetMigration_t}{AveragePop_t}) * 1000$

[46] $AveragePop_t = (\frac{Population_{Jan1,t} + Population_{Dec31,t}}{2})$

2.4.2 Micro-data: administrative individual-level data

Administrative data for the Netherlands is obtained from Statistics Netherlands[47] in which individuals are uniquely identified by a Random Identification Number (RIN). Due to the nature of the data, two samples are set up. First, an unrestricted sample in which individuals are observed if they are alive and registered in a Dutch municipality by 1995. Second, a restricted sample in which individuals are observed if the individual and its mother are alive and registered in a Dutch municipality by 1995. The unrestricted sample is larger, but more limited in the availability of information on household characteristics as compared to the restricted sample. The restricted sample contains a different set of individuals (selective sample), which is exploited to study sample selection.

A conceptual framework for 'culling'

This paper focuses on labor market and health outcomes in adulthood (I observe labor market outcomes when the studied cohorts are aged 51 to 55). Any study that examines the long-run effects of historical events, and especially those that occur before birth, copes with selectivity of the survivor population. This paper considers selection into parenthood, but there may be additional selectivity of the survivors, i.e. which children are born alive, and survive until adulthood. This is often referred to as 'culling'. Culling is important when studying the long-term effects of pandemics and disease (e.g Almond, 2006; Mamelund, 2006), as the survivors are probably

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

in better health. Negative parental selection may also lower the child's odds of survival onto later ages. The interaction between parental selection and mortality can be formalized in a latent variable framework.

Let h_i^* denote the unobserved characteristics of individual i (following Almond, 2006). The individual's unobserved characteristics h_i^* are a function of their own unobserved characteristics (e.g. genes) c_i^*, and parental quality p_j.

$$h_i^* = f(c_i^*, p_j) \tag{2.3}$$

Higher levels of h_i^* are associated with larger odds of survival to later ages, up to the point that individuals are observed in the data (denoted by T). If h_i^* falls below a threshold level d, the individual does not survive until adulthood. The mortality rate prior to adulthood (before T), denoted by Early Mortality Rate (EMR), can be defined by the cumulative distribution function $F(h_i^*)$:

$$EMR \equiv F(d). \tag{2.4}$$

The survival rate (SR) for individual i at time T, i.e. those that have survived until adulthood, is then given by:

$$SR \equiv 1 - F(d). \tag{2.5}$$

I assume that early life mortality (i.e. before T) is most likely to occur for those individuals with lower h_i^*. Implying that, conditional on having survived until adulthood and by assuming that lower h_i^* coincides with lower ability, their outcomes would have been worse compared to the indivi-

duals who are observed in the sample.[48] Due to this selection, the estimated effects of parental selection are biased upwards, i.e. be less negative in case of negative parental selection or more positive in case of positive parental selection.

The selectivity is more complex for the restricted sample, when conditioning on the survival of the mother until year T. Let p_j denote parental characteristics of mother j, which are a function of unobserved parental quality q_j^*, and maternal age at birth a_j.

$$p_j = g(q_j^*, a_j) \tag{2.6}$$

Parental quality is a proxy for unobserved factors contributing to the mother's survival until T. Age at birth mechanically lowers the odds of survival as the probability of death (before T) increases with age at birth. Higher levels of p_j are associated with larger odds of survival. p_j is increasing in parental quality q_j^* and decreasing in age at parenthood a_j. Hence:

$$\frac{\partial p_j}{\partial q_j^*} > 0 \quad \text{and} \quad \frac{\partial p_j}{\partial a_j} < 0. \tag{2.7}$$

Parents die before year T if p_j falls below threshold level e. The individuals who are alive by T and whose mother is alive by T are different from those whose mother has died. They can be different as (1) lower quality parents died in which case average h_i^* will be higher, (2) older parents died, and when assuming that these parents have better characteristics, h_i^*

[48] With the data I have it is not possible to examine the selectivity in the survivor population. However, section 2.6.3 discusses insights from other sources that can shed more light on the characteristics of the survivor population.

average would be lower.

Unrestricted sample

For setting up the unrestricted sample I start with the Municipal Population dataset (Gemeentelijke Basisadministratie 2016), which contains data of all individuals registered in Dutch municipalities by 1995. Available variables include the individual's birth date, birth dates of parents, gender, and country of origin. The sample is restricted to individuals born between 1944 and 1948 (N=1, 161, 250). Individuals born outside of the Netherlands (N=120, 338) are dropped from the sample. The data is merged with the place of birth dataset. The file contains information on place of birth for individuals registered in municipalities by 2004. Individuals for whom place of birth cannot be determined (N=105) and individuals born in Flevoland[49] (N=1, 168) are dropped from the sample, as well as individuals whose mother's age is missing (N=926). The resulting sample contains 530, 113 males and 508, 600 females (see Table 2.1).

This paper exploits geographical differences in the magnitude of the Birth Peak. The different municipalities are assigned to provinces using municipality codes.[50] Children are born in the south if their place of birth is in the provinces of Limburg, Noord-Brabant, or Zeeland. North contains children born in Zuid-Holland, Noord-Holland, Utrecht, Groningen, Friesland, Drenthe, Overijssel, and Gelderland. All children born in the large cities (more than 40, 000 inhabitants in 1944) in the west are excluded as these were affected by the famine. Following Stein et al. (1975)

[49]Flevoland is the youngest province of the Netherlands (dating back to 1986), most of its land was reclaimed in the 1950s and 1960s after a flood in the early 1900s. Hence the population in this area was very low in 1946.

[50]Available at statline.cbs.nl, *Gebieden: overzicht vanaf 1830.*

and Scholte et al. (2015) this excludes children born in Amsterdam, Delft, Haarlem, Leiden, Rotterdam, the Hague, and Utrecht.

Restricted sample

Children are matched to their parents to get more information on family and parental characteristics. The Parent-Child dataset matches children to parents who are alive and living in the Netherlands between 1995 and 2015, and contains information on 15,860,240 individuals. Children can be matched if their mother is still alive by 1995.[51] Stillbirths (N=22,290) and individuals whose mother's RIN is missing (N=547,350) are dropped from the sample. This data is supplemented with demographic characteristics from the Municipal Population dataset (in Dutch: Gemeentelijke Basisadministratie). A sample of 15,284,285 individuals remains after merging the demographic data to the Parent-Child dataset. In the remaining sample children are defined as siblings if they share the same mother. Birth order is defined based on the child's date of birth relative to siblings. Individuals whose mother's age at birth cannot be determined (N=3,912) are dropped from the sample. The sample is restricted to individuals born between 1944 and 1948 (N=735,085). Individuals who are not born in the Netherlands (N=23,086) are dropped from the sample. Data on the individuals' place of birth is added. Individuals whose place of birth cannot be determined (N=25) are dropped, as well as individuals reported to be born in Flevoland (N=829). The remaining (restricted) sample contains 363,023 males and 348,122 females and is shown in Table 2.1.

Information on the marital state of the parents during childbirth (from

[51] Table 2.A2 shows maternal age at birth in 1946 and the respective age of the mother in 1995.

gbaburgelijkestaatbus) is added. The dataset contains all past and present marital statuses (unmarried, married, registered partnership, etc.) for individuals registered in Dutch municipalities from 1995. From the individuals in the restricted sample 100,082 cannot be matched to a marital status. The remaining sample contains 310,240 males and 300,823 females. Although 14.07% of observations cannot be matched to information on the marital status, the observed percentage of in-wedlock births in the data in 1946 (i.e. 97.3%) closely reflects the actual population average of 97.5%. In-wedlock conceptions are defined conservatively as childbirth within seven months after marriage to account for the possibility of premature births. Shotgun marriages make up to remaining share of in-wedlock births, and represent a situation in which a child is conceived out-wedlock and born in-wedlock. They portray a situation in which the child is conceived before a couple is married, and likely reflect unanticipated conceptions given the nature of family formation at the time.

Labor market and health outcomes

Data on labor market outcomes is available starting from 1999. This paper focuses on labor market outcomes in this particular year, when the studied cohorts are aged 51 to 55, to capture as many individuals as possible in active labor market positions.[52] This is especially important as the cohorts had access to an early retirement scheme, which is not yet available for the studied cohorts in 1999 (Scholte et al., 2015). Information for labor earnings is retrieved from the dataset with yearly taxed earnings from all paid employments in 1999 (*baanprsjaarbedragtab*), and from the dataset with all

[52]It is also the first year in which these outcomes are observed in the administrative data from Statistics Netherlands.

earnings from self-employment in 1999 (*zelfstandigentab*).[53] An indicator for labor market participation is created which is equal to one when labor earnings from the two sources mentioned above are greater than zero in 1999. An indicator of the individual's enrollment in any social security scheme is created with data on whether the individual received benefits from disability insurance (*aototpersoonsbus*)[54], unemployment insurance (*wwjaarbedragtab*), or welfare (*bijstandjaarbedragtab*) in 1999.

The analysis on labor market outcomes is restricted to men due to the nature of female labor force participation for the studied cohorts. First, until the 1960s women were not allowed to work after marriage, implying that labor market outcomes may well reflect marriage market success. Although after 1957 women in public service jobs could no longer be fired, it was not immediately acceptable for women to work, especially after having children (De Graaf and Keil, 2001).[55] Second, the cohorts studied in this paper show a steep increase in labor force participation at late ages (Arts and Otten, 2013). Comparing the late-life labor market outcomes of these cohorts of females may pick up this general trend in labor force participation.[56]

The analysis on health outcomes, i.e. mortality and the use of prescription drugs, is done for the sample of men and women. Information

[53]Income from self-employment is negative for 0.85% of the sample, those observations are recoded as having zero income. One cent is added for observations with zero income after which the log is taken. The results for estimating the models conditional on having positive earnings, hence leaving out observations with zero or negative earnings are shown in Table 2.A5.

[54]Including the general scheme for employees (WAO), the scheme for young individuals (WAJONG), and the scheme for self-employed individuals (WAZ).

[55]To illustrate, only 3 out of 10 women at the start of the 1980s had a paid job for more than 12 hours a week (De Graaf and Keil, 2001).

[56]This is one of the reasons, apart from the famine potentially affecting reproductive capacities, why Scholte et al. (2015) do not consider female labor market outcomes.

on mortality is available in the dataset with all dates of death of individuals registered in Dutch municipalities by 1995 (*gbaoverlijdentab*). Implying that the mortality measures only capture individuals who have survived up until 1995 and died afterwards (i.e. survived until age 49 for those born in 1946). Indicator variables are created that represent mortality before age 65 and before age 70. The latest data-file is available for 2017, and for this reason I cannot capture mortality beyond age 70 yet. The medication file (*medicijntab*) contains information on prescription drugs (ordered by ATC-4 code) provided and covered by basic health insurance in 2006.[57] I focus on prescription drugs associated with diseases that could potentially arise due to growing up in a different family environment. This includes prescription drugs for mental health problems (anti-psychotics, anxiolytics, sedatives, and antidepressants),[58] cardiovascular diseases (including hypertension), diabetes mellitus, and respiratory diseases.[59] Unfortunately the data does not show how many prescriptions the individual received in each category in a year, but rather whether the individual received any prescription in the specific ATC-category in a given year. An indicator is created which is equal to one if the individuals received any prescription in 2006 in the four categories discussed before. Models with indicators for receiving prescriptions drugs in the four separate categories, as well as an indicator for having received any prescription drugs (irrespective of whether drugs could be related to diseases caused by lifestyle) are also reported.

[57]Basic health insurance is a mandatory insurance for all inhabitants of the Netherlands that covers necessary care.

[58]These correspond to respectively ATC-codes N05A, N05B, N05C, and N06A.

[59]There is no data available for earlier years. Following Huber et al. (2013) an individual is classified as having prescription drugs for cardiovascular diseases if he or she takes drugs with ATC-codes B01A, C01, C02, C04A, C07, C08 and C09. For diabetes mellitus this includes drugs with ATC-codes A10A, A10B, and A10X, and for respiratory diseases this includes drugs with ATC-code R03.

Table 3.4 shows descriptive statistics for the sample discussed. Parents are on average slightly older in the south, and family size is also bigger in the south. As all big cities in the west of the Netherlands are left out the analysis due to the famine, the urban indicator (equal to one for cities with more than 40,000 inhabitants in 1945) is lower in the north versus the south. Mortality after 1995 is very similar in north and south for the studied cohorts, although the use of prescription drugs that could be related to growing up in a particular family is higher in the south. Labor market outcomes are better for children born in the north, as they have higher employment, higher earnings, and a lower enrollment in social insurance. Table 2.3 focuses on first births and splits the sample by in- and out-wedlock conceptions. Children born in shotgun marriages have younger parents, and more siblings. What stands out is that shotgun conceptions do worse than in-wedlock conceptions on the majority of outcome variables. Particularly, they have higher mortality before age 70, higher use of prescription drugs related to lifestyle, lower employment, lower earnings, and a higher enrollment in social insurance.

2.5 Results

2.5.1 Family environment

This paper argues that socioeconomic conditions may change parental selection and hence the family environment in which a child grows up. Family environment again affects the child's human capital accumulation, which can have consequences that extend towards adulthood (e.g. Cunha and Heckman, 2007; Björklund and Salvanes, 2011). This section examines

outcomes related to family environment and hence can be interpreted as a first stage.

Table 2.4 shows the difference-in-difference estimates with two family characteristics as outcome variables, namely the number of children in a family, and the stability (length) of parental marriage. The CoBP grow up in families that are significantly smaller by 6.8% compared to cohorts born before and those born in the south. This finding makes sense if parents delayed fertility, as there is simply less time (biologically) to reproduce when starting at a later age. Similarly it might be that if the first birth is unanticipated, parents may have less subsequent children. When focusing on first births and by making a distinction between delayed conceptions and unanticipated conceptions, it seems that the lower family size effect is driven by planned births. Hence, parents who conceived their child in-wedlock and who are more likely to have delayed fertility, have lower completed fertility. Note that this model only compares children who are conceived and born in-wedlock, suggesting that the planned births in the 1946 cohort are different from those in earlier cohorts. The point-estimate is however less precisely estimated which decreases significance, but the size of the coefficients increases (-0.274 vs -0.311). There is no evidence for an effect on family size for unanticipated conceptions, i.e. the point-estimate is much smaller (-0.037) and very imprecisely estimated. Suggesting that within the group of unanticipated conceptions, those born in 1946 in the north do not grow up in families of different size as opposed to earlier cohorts.

When considering the length of parental marriage as an outcome variable, the difference-in-difference is not significantly different from zero.

However, when only regarding children who are born in-wedlock, and making a distinction between children who are conceived in wedlock and those who are not (shotgun marriage), an interesting picture emerges. The difference-in-difference estimator is not significantly different from zero for children who are conceived in-wedlock. However, the length of parental marriage is significantly shorter for CoBP conceived in shotgun marriages. More precisely, all else equal the marriage of their parents takes on average 5.6% shorter.[60] CoBP born from unanticipated pregnancies grow up in less stable households, which may reflect lower quality match of their parents, which is suggestive evidence for negative parental selection.

2.5.2 Labor market outcomes

The results for labor market outcomes are shown in Table 2.5, and remember that these models are only estimated for males. Panel A reports the results from the difference-in-difference model where children born before and after the liberation are compared in the north and south of the Netherlands. First, note that the point-estimates change with the inclusion of controls. As the outcomes are measured at one point in time (i.e. 1999) and the cohorts are of different age at that particular point, the specification controls for these age profiles by including linear and quadratic region-specific age trends. This, together with controls for being conceived during the liberation of the south and the inclusion of birth month dummies that control for birth circumstances, and a control for being born in the city to additionally control for regional differences, makes that point estimates change with the inclusion of controls. The dummy for being

[60] Using a two-sample t-test I find that the difference between the difference-in-difference estimators of in- and out-wedlock conceptions is significant with $P < 0.05$.

born from March to September 1946 is associated with better labor market outcomes in the specification without controls, whereas the coefficient gets smaller and insignificant with the inclusion of controls. Suggesting that age-trends do indeed matter for labor market outcomes. Another pattern that appears is that children born in the north have significantly better labor market outcomes, which may be caused by differences in occupations across regions.

When looking at the difference-in-difference estimator (i.e. $CoBP$), no clear picture emerges from the table when considering various labor market outcomes. Implying that on average, there is no evidence that CoBP fare significantly better or worse as compared to those conceived before the liberation. If anything, although imprecisely estimated, the point estimates suggest that they would do worse on all three labor market outcomes considered (lower employment, lower earnings, and higher enrollment in any social insurance program). Notice that the point-estimates for log labor earnings are particularly large (but estimated very imprecisely). CoBP would on average have 11% lower earnings. However, the zeros for the unemployed are included in this outcome measure by adding up one cent to earnings before taking the log. When considering a conditional earnings measure (see Table 2.A5) which conditions on employment, another picture emerges. Conditional on being employed, CoBP would earn 2.6% more, although imprecisely estimated. Hence it seems that the large negative point-estimate in Table 2.5 is driven by the zeros in the outcome variable.

The absence of an effect in this aggregate approach might be caused by the fact that two types of behaviors led to births in 1946, namely delayed

fertility and unanticipated conceptions. The aggregate approach may mask these two potentially offsetting effects. Panel B endeavors to distinguish above-mentioned behaviors by restricting the attention to first births who were born in-wedlock, and by making a distinction between children who were conceived in-wedlock and those who were not (i.e. born in shotgun marriages). The first are likely to be a product of delayed fertility, whereas the latter likely reflect unanticipated pregnancies. Remember that the CoBP coefficient should be interpreted within the group considered. That is, for unanticipated conceptions it measures how shotgun conceptions conceived after the liberation of the north are different from those born before the liberation and in the south. Interestingly, also within each category there is a premium of being born in the north as those children have better labor market outcomes. However, this coefficient is only significant for in-wedlock conceptions which may be caused by the lower sample size and consequently lower precision for the shotgun sample. Although the difference-in-difference estimators are not significant for any of the labor market outcomes considered, it is interesting to examine the sizes of the coefficients in both groups. CoBP conceived in-wedlock have lower employment, lower earnings, and lower enrollment in any social security system, whereas CoBP born in shotgun marriages have lower employment, higher earnings, and lower enrollment in social insurance. Summarizing, Table 2.5 does not provide clear evidence that CoBP fare better or worse. Similarly, when a distinction is made between planned and unplanned births no clear pattern emerges.

2.5.3 Health outcomes

The results for the health outcomes are reported in Table 2.6. The models for health outcomes are estimated for both genders combined, which explains the about doubling of the sample size compared to the analysis on labor market outcomes.[61] Panel A shows the results of the baseline difference-in-difference model. Note that the inclusion of controls does not have a big effect on the mortality indicators, which is because they are measured at a fixed age (i.e. before age 65 and 70). Including controls does affect the point estimates for having had any prescription drugs in the four relevant categories (i.e. drugs for mental health, respiratory diseases, cardiovascular diseases, and diabetes), as this outcome is measured in 2006. The results for mortality in Table 2.6 are not statistically significant, and also economically very small. When regarding the use of any prescription drugs the point-estimate is negative and significant. CoBP have a 2.4 percentage points lower use of prescription drugs in the four relevant categories. When distinguishing between the four drug groups in Table 2.A6 it seems that this negative effect is driven by lower prescriptions (-1.8 percentage points) for mental health drugs for CoBP.[62]

The difference-in-difference estimator on the aggregate level may mask heterogeneity among planned and unplanned births. Panel B focuses on first births and distinguishes between those who were conceived in-wedlock and those born in shotgun marriages. Although imprecisely estimated, the

[61] The results are robust to estimating the models for both genders separately (see Table 2.8).

[62] Table 2.A6 also includes an outcome variable for the use of any drugs, irrespective of whether the drug may be related to lifestyle and family environment. Although a significant point-estimate is found for this outcome, it should not receive much attention. This outcome variable also includes drugs for coincidental and temporary diseases and hence does not tell us much.

effects on mortality before 65 seem larger among unplanned births, whereas the effects on mortality before age 70 are very similar in both groups. The effects on prescription drugs is larger among unplanned births, although estimated imprecisely. When considering the results in Table 2.A6 both effects seem again driven by an lower use of mental health drugs. To sum up, the results in Table 2.6 do not show strong evidence that CoBP fare better or worse. They do have a lower use of prescription drugs which are relevant to family environment, and particularly prescription drugs related to mental health problems.

2.5.4 Robustness checks

The results show very little evidence for the CoBP being different in terms of family environment as well as later life labor market and health outcomes. This section explores the robustness of the results. First, the main analysis compares children before and after the liberation in north and south as these children are exposed to different socioeconomic circumstances. A robustness check is done in which the Birth Peak cohort (born March-September 1946) is compared to later cohorts (i.e. born from October 1946 to December 1948). Both cohorts are exposed to better socioeconomic conditions. Table 2.7 and 2.8 show the robustness to using later cohorts as control group on respectively the family and child level.[63] On the family level, the finding on family size disappears, suggesting that CoBP do not have smaller families when comparing with cohorts born after. The results on the length of parental marriage and labor market outcomes show very similar patterns. The point-estimate for the use of prescription drugs in

[63]Full estimation results are available in Table 2.A7, Table 2.A8 and Table 2.A9.

the four relevant drug groups is no longer significant (which is caused by a decrease in the point-estimate). Overall, the main conclusions do not change when using the post-cohorts.

Second, I check the robustness of the results to using different sub-groups. The main results do not condition on birth order, whereas the sub-sample analyses conditions on first births who are born in-wedlock. The analysis is redone for only first births and the results for family and child outcomes show a similar pattern. The analysis on health outcomes is estimated for a joint sample of men and women. The results are robust to estimating the models separately by gender. Then, the main model on child outcomes is estimated for out-wedlock births only. Notice that this is a very small sample due to the nature of family formation at that time. Within this sample of illegitimate births, those conceived in times with better socioeconomic do not fare better or worse in adulthood.

Third, the main analysis compares children born in the north and the south of the Netherlands. One could argue that these areas are very different as the outer regions of these areas are geographically dispersed. To improve comparability I re-estimate the main model by focusing on provinces in the north and south who are closer geographically (i.e. comparing the south with the provinces of Zuid-Holland, Utrecht and Gelderland). The results are reported in Table 2.7 and Table 2.8. Again the results are very similar to the main estimates in this robustness check. Only the coefficient for the use of prescription drugs decreases and loses significance, the coefficient on the parents' marital stability and mortality before 70 increase and become marginally significant.

Fourth, I investigate the robustness to different definitions of treatment

and control group. The BP-cohort is defined as all birth between March and September 1946, as the number of births started to increase from March 1946. These children are, when assuming a pregnancy duration of nine months, conceived from June to December 1945. The liberation occurred in May 1945, implying that children born in February may also be a part of the Birth Peak cohort. The Birth Peak cohort is redefined as births from February to September 1946, and the results show a similar pattern. Two differences are that the coefficient of interest for employment increases and becomes marginally significant, and the point-estimate for the use of prescription drugs decreases and is no longer significantly different from zero. Another robustness check is executed in which the Birth Peak cohort is redefined as born between March and May 1946, which are the months in which birth rates were highest. These could be the months in which conceptions were mostly driven by the changed socioeconomic conditions. The results are robust to using this different definition as point-estimates are very similar for the considered outcomes.

As a next step I examine the robustness to leaving out children who were conceived in periods characterized by war circumstances that could also have affected on fertility. The main analysis leaves out children born in famine-affected areas, which are defined as the large cities in the western part of the Netherlands. As a robustness check I leave out children born between March and December 1945, who were probably conceived during the famine. The results for family characteristics, and child outcomes are robust to using this different definition. The one difference being that the coefficient on family size drops from -0.309 to -0.185. I also check the robustness of the results to leaving out those who are conceived in the time

period that the south was liberated, but the north was still occupied. The main conclusions do not change when this definition is chosen, except the point estimate for the use of any prescription drugs which decreases in size and is no longer significant.

Finally, to investigate potential selection issues for the relevant outcome variables I condition on mortality for the outcomes considered. That is, for child outcomes, I condition that the individual must be alive in 1999 (for labor market outcomes) and 2006 (for health outcomes). The results in this robustness check show a similar pattern, it only mechanically changes the point-estimates for the mortality indicators.

2.6 Mechanisms at work

So far I find no effects of improved socioeconomic conditions on parental selection nor long-term child outcomes. Even though the change in socioeconomic environment studied in this paper can be considered as a severe shock, parental selection effects to economic uncertainty to do not translate to this particular setting. It could be that economic uncertainty has different effects on the decision to have a child (as discussed in Gronau, 1977) as opposed to the end of war. Hence, the income and substitution effects that occur as a response to economic uncertainty do not extrapolate to a context where socioeconomic conditions improved due to the end of war.

Similarly family size does not seem to affect long-term child outcomes in my estimations. The quantity-quality trade-off of family formation predicts that family size is negatively linked to child later life outcomes (e.g. Becker

et al., 1973), although Black et al. (2005) find no causal effect of family size on later life outcomes in Norway, which is in line with the findings in this paper. Moreover, in a sub-sample for which I do find parental selection, in terms of a decreased household stability, I find no persistent effects until adulthood. This is in contrast with the literature which generally finds large effects of family environment on child long-term outcomes (e.g. Cunha and Heckman, 2007; Björklund and Salvanes, 2011). However, my findings do not invalidate the findings in other studies. First, note that this paper focuses on the effect of parents' marital stability on outcomes that persist for over fifty years. Suggesting that although the length of parental marriage does not seem to be related to very long-term child outcomes, this does not mean that there is no effect of family environment on child outcomes. Second, evidence shows that family environment is less important in the Nordic countries as opposed to the United States. Likewise, it appears that changes in educational policy can also influence the effect of family environment on child long-term outcomes (Black et al., 2011). The institutional setting studied in this paper is more related to the Nordic setting, suggesting that inter-generational persistence may be less strong in the Netherlands. A high equality of opportunity may reduce the persistent effects of family environment on child outcomes.

The next sections explore three additional mechanisms that may explain the (lack of) effects found. First, it it shows that cohort effects cannot explain the main findings in this paper. Second, it focuses on the role of war circumstances, apart from changed socioeconomic conditions, that may have affected fertility and parental selection right after the war. More precisely, it concentrates on the role of Hunger Winter, the absence of men,

and the influence of the Canadian liberators. It seems that the changed socioeconomic environment is indeed the largest contributor to the fertility rise. Finally, I consider the role of sample selection effects, as it could be that potential adverse effects cannot be identified as they may have occurred earlier in the life-cycle.

2.6.1 Cohort effects or parental selection?

It could be that any parental selection effects are offset by cohort-specific effects. The 1946 cohort may not only be different in composition but is for sure different in size. The cohort experienced large classes in school, but the entrance onto the labor market for this cohort coincided with a large demand for labor. Overall, being born in this particular cohort could influence labor market and health outcomes through other ways than parental selection, e.g. through the size of the cohort. That cohort size matters is also shown in the literature (e.g. Bound and Turner, 2007; Brunello, 2010), and potential positive parental selection can be offset by negative cohort-specific effects. I perform two tests to check whether cohort size or cohort composition is driving the results. First, I add controls for the size of the child's birth cohort at the province level to the main specification. The province level is chosen to capture potential crowding effects in schools or on the labor market. As of the low residential mobility of the Dutch it likely represents a good proxy. The results are shown in Table 2.12.[64] The point estimates are very similar when including these cohort level controls, suggesting that cohort size does not contribute to the effects found.

A stronger test is performed in which I examine the outcomes of younger

[64]Full estimation results are available in Table 2.A10 and Table 2.A11.

siblings.[65] Specifically, I select all children born between January 1947 and December 1950 in families in which at least two children were born between March 1946 (start of the Birth Peak) and December 1950. The analysis is restricted to these families as having a second (or further) birth may be endogenous to earlier births, and these families are similar as they choose to give birth to at least two children after the liberation. I also estimate the models for a group of families that gave birth to exactly two children after the liberation to address the endogeneity issue. Children born in this time-frame who have an older brother or sister born during the Birth Peak in the north are classified as *CoBP siblings*. These children were not affected by the same cohort-specific factors as their older brother and sisters, but they are born to the same parents. Implying that any cohort effects that may have played a role for the BP cohort is not important for these children. If the *CoBP siblings* are affected in a similar way as their older brothers and sisters, one could argue for parental selection. If not, it is likely that cohort effects instead of parental selection are driving the results. A model analogous to the main difference-in-difference approach used in this paper is estimated for this particular sample. Yet this time the coefficient of interest is an indicator for having a CoBP sibling.

The results for labor market outcomes are shown in Table 2.9. Panel A shows that the estimator of interest (*CoBP siblings*) is not significantly different from zero for any of the labor market outcomes considered. Implying that there is no evidence for CoBP to be different, and a similar picture emerges for their siblings. If anything, the coefficient of interest is positive (although estimated imprecisely). Panel B separates the sample by

[65] A similar strategy to separate parental selection from bad circumstances is executed by Chevalier and Marie (2017).

marital state at first birth. Implying that it distinguishes between children whose first born brother or sister was conceived in-wedlock or conceived in a shotgun marriage. Again none of the difference-in-difference estimators is significantly different from zero.

Table 2.10 reports the results for health outcomes. Panel A shows that the coefficient of interest is not significant for the health outcomes. Suggesting that also for these health outcomes, there is no evidence for cohort effects instead of parental selection are driving the results found in this paper. Panel B separates the sample by marital state at first conception (i.e. first birth was conceived in-wedlock or in a shotgun marriage). Again in the coefficient of interest is not significant for any of the outcome variables considered. Some of the point-estimates change, but as they are estimated imprecisely it does not change the conclusion. Overall, as the coefficients of interest are not significantly different from zero for both the CoBP and CoBP siblings, I cannot conclude that cohort effects are driving the effects found.

2.6.2 The influence of other war-related explanations

Apart from changed socioeconomic conditions, Section 2.2.4 discussed several other explanations for the Birth Peak. This section explores how much these explanations contribute to the effects found. Based on the tests provided below, it seems that the change in socioeconomic environment can be considered as the largest contributor to the fertility rise.

The influence of the Hunger Winter

As discussed in Section 2.2.4, the Hunger Winter could have affected fertility, and consequently parental selection. To avoid that the famine was affecting the results, children born in famine-affected areas (the large western cities) are left out of the treatment group. Equivalently, a robustness check was done in Section 2.5.4 dropping all children who were conceived during the famine, which showed very similar results. To explore the influence of the famine further, I do the analysis including children born in famine-affected areas (Table 2.12). The point estimates are very similar to those in the baseline model. Summarizing, the above mentioned checks suggest that the results are not sensitive to the influence of the famine that occurred in the winter prior to the liberation.

The absence of men

Another explanation for the Birth Peak as discussed before, is the absence of men during war years. Likewise, fertility can increase after the war with the return of these men. According to my estimates, at its highest point about 31% of men in childbearing ages were in forced labor. However, even during the war fertility was increasing and responding to war circumstances suggesting that the absence of men cannot be the primary explanation for the fertility rise. The largest losses of men to forced labor came at the end of war (fall of 1944) when raids took place in Rotterdam and The Hague were taken abroad. The raids in Rotterdam alone accounted for about 14% of all men in forced labor.[66] As Rotterdam and The Hague

[66] Author's estimates based on the approximation that about 50,000 men were taken away in November 1944, on a total of 351,000 men in forced labor (estimate from Krimp and Kemperman, 2015)

are left out of the treatment group as of the influence of the famine in these cities, this would also partially address the disproportional retrieval of men out of these municipalities. To the best of my knowledge there is no information on the actual location of these men in Germany. Hence, I cannot exploit differences in distance or difference in when these men were allowed to go back to the Netherlands. However, I can estimate the model excluding children born in provinces bordering Germany, and assuming that men could return more easily/quickly when living in these bordering provinces. Although the Netherlands is small, and this test is imperfect it may provide some suggestive evidence. Table 2.12 shows that the results are very similar to the main estimates when dropping children born in border provinces. Based on this test, the influence of the absence of men due to forced labor on parental selection seems limited.

The influence of the Canadian liberators

It could be that the presence of the Canadian soldiers led to extra births, and that any parental selection is also driven by the presence of these soldiers. Although it is unlikely that the Canadian liberators had a major impact on fertility at the time - a conservative estimate shows that only 1.2% of all babies born in 1946 are born to Canadian liberators (see Section 2.2.4 for more detail) - I examine the robustness of the results to excluding births that were more likely to be from these Canadians soldiers. Two strategies are employed: (1) one that exploits the timing of the exit of the Canadian forces out of the Netherlands, and (2) one thats exploits the location of the Canadian soldiers across the Netherlands.

The Canadian soldiers started leaving the Netherlands from June 1945

(17,000 soldiers), July (26,000 soldiers), October (110,000 soldiers), and the last soldiers left near the end of May 1946 (Bollen and Vroemen, 1994). As the majority left before October 1945, children born from August 1946 are less likely to be from the Canadian liberators. I redefine the treatment group as being born from August-September 1946. Those born from March-July are temporarily dropped from the sample. The results are reported in Table 2.12. Although some point estimates change, which is to be expected when leaving out the children that were conceived in the months right after the liberation, leaving out these children does not alter the conclusions.

The liberation of the northern provinces failed in the fall of 1944 and the Allied forces had to re-group to resume the liberation attempt of the north. Many Canadian soldiers had to be housed in the Netherlands, and had to be entertained (Hofstee, 2012). So-called leave-centers were set-up (first only in the south, but after May 1945 also in the north). Soldiers were allowed to bring a 'plus one' back to the leave-centers, which led to the nickname 'love-centers' (Okkema, 2012). This is where a lot of the 'engagements' took place, and it is likely that girls living in the neighborhood of these leave-centers had a higher probability of engaging with the Allied forces.[67] I drop children born in cities with such a leave-center and re-estimate the difference-in-difference model, the results are reported in Table 2.12. The point estimates are similar to those in the baseline model and hence the conclusions do not change. Summarizing, from the two

[67]Before the liberation of the North these centers were located in Nijmegen, Breda, Tilburg, Eindhoven, Den Bosch, Heerlen and Maastricht. After the liberation of the north they were stationed in Amsterdam, Almelo, Arnhem, Deventer, Enschede, Nijmegen, Zwolle, Amersfoort, Apeldoorn, Barneveld, Hilversum, Utrecht, Asssen, Groningen, Harlingen and Leeuwarden (Okkema, 2012). Kleinhout (2006) argues that the leave-center in Amsterdam was most important, followed in popularity by Utrecht, Groningen, Apeldoorn and Enschede.

checks in this section I cannot conclude that the presence of the Canadian soldiers is driving the findings in this paper.

2.6.3 Sample selection effects

Last but not least, it could be that sample selection effects drive the estimate to zero. Section 2.4.2 discussed 'culling' of the weakest, which might occur when late-life outcomes are studied. Specifically, the individuals who are observed in the data are hypothesized to have better characteristics than those who are not. This is particularly important as the individuals who were most negatively affected/selected, may die prematurely and may thereby be no longer observed in the data. To illustrate, Van Ewijk and Lindeboom (2017) find no long-term health effects of prenatal exposure to WWII, if anything health is better, which they explain among other things with selective mortality. This section addresses potential selectivity into the sample.

First, parents with different characteristics may *decide* to conceive in different times, but it could also be that different parents are *capable* of conceiving in different times. Section 2.2.5 established that parental selection is not driven by available care. However, parental selection may be affected by mortality of men and women in childbearing age during WWII. If men and women who died are different from those who did not, children born in 1946 might be born to an even more selective set of parents. Figure 2.A7 shows mortality in the Netherlands from 1936 to 2015. Mortality during the war is especially high for individuals aged 45 to 80. In the group of men and women in childbearing age (aged 15 to 45) a peak of 7 deaths per 1000 inhabitants is observed in 1945. Mortality is even lower in other war

years, and although there is no information available on selective mortality, the low mortality in the relevant age group suggests that it will only play a minor role. Second, there could be selection at birth. That is, conditional on getting pregnant there may be differences in which women brought their pregnancies to term. Induced abortion was not legal in 1946, implying that (legal) selective abortions will not affect which children are born. Likewise there are no trend-deviations in the number of stillbirths and perinatal deaths for 1946 (Figure 2.A6), and the sex ratio is more skewed to boys in 1946.[68] Unfortunately, there is no data available on these outcomes on the regional level.

The main analysis uses the restricted sample in which more information on family characteristics is available. Individuals are observed in the restricted sample if they themselves and their mother are still alive and registered in a Dutch municipality by 1995. The restricted sample contains 166,475 individuals born in 1946, which is 58.5% of the total amount of births. The differences between the unrestricted and restricted sample were formalized in Section 2.4.2 and two processes underlie the survival from the unrestricted to the restricted sample: (1) individuals with a lower quality mother are less likely to be observed, and (2) individuals whose mother is older at giving birth have a lower probability of being observed. Table 2.11 shows the results from the difference-in-difference models for the unrestricted sample with an indicator for observing the mother in the restricted sample as the relevant outcome. As predicted, a higher maternal age at birth is associated with a lower probability of observing the moth-

[68] According to the fragile male hypothesis this does not raise concerns on bad pregnancy circumstances Figure 2.A8. Bethmann and Kvasnicka (2014) argues that in many European countries, including the Netherlands, the sex ratios during and right after WWII were skewed towards boys.

er. The double-difference estimator, that captures the effect of being born March-September 1946 in the north, is not significantly different from zero for both the pre- and post-sample. CoBP are not more or less likely to be observed, which suggests that both processes outlined in Section 2.4.2 balance each other. In line with this finding, when estimating the difference-in-difference model for the unrestricted sample (see Table 2.12) a similar pattern arises, suggesting that there are no large differences between the unrestricted and restricted sample.

The total number of births was 284, 456 in 1946, and 242, 196 of those births are observed in the unrestricted sample (alive and registered in a Dutch municipality by 1995), which is about 85.1% of the total number of births. The discrepancy can arise from both deaths and migration. It may be that children who are most negatively affected die before reaching adulthood, and thereby are not observed in the data. Lindeboom and Van Ewijk (2015) find that lower survival probabilities until age 55 for those born directly after the war. However, this relationship disappears after conditioning on survival up to age one and five. That lower survival probabilities are probably caused by higher child mortality, is consistent with Figure 2.A6 that shows that mortality after birth is higher for the 1946 cohort when comparing to later cohorts. Rau et al. (2017) argue that an increase in unanticipated conceptions caused by a large price increase in the price of contraceptives led to higher infant mortality, which could be coherent with negative parental selection. Similarly, Gruber et al. (1999) find that access to abortion, a measure that can prevent unanticipated births is associated with lower infant mortality. Unfortunately such potential effects earlier in the life-cycle cannot be studied with the current data.

2.7 Conclusion

This paper examines how the socioeconomic environment at conception influences the size and composition of a cohort. It specifically explores whether changed socioeconomic conditions, induced by the end of war in the Netherlands, affect parental selection and consequently child outcomes in adulthood. I exploit a natural experiment as regional differences in the timing and magnitude of the changed socioeconomic conditions led to a Birth Peak in the area that experienced the most severe change in socioeconomic environment. The timing and regional differences in the magnitude of changed socioeconomic conditions are exploited in a difference-in-difference framework.

In my analysis I find no evidence of parental selection, as measured by the stability of the parental marriage, nor do I find effects on long-term child labor market and health outcomes. This implies that the parental selection effects that are documented to occur with changes in economic conditions do not extrapolate to the context studied in this paper. It could be that the forces behind the demand for children respond differently to changed economic conditions than to changed socioeconomic conditions caused by the end of war.

The absence of an effect found could be explained by heterogeneity between conceptions arising due to delayed fertility and unanticipated conceptions. Different types of parental selection may underlie these two types, which could net out on average. When distinguishing between planned and unplanned conceptions, I find that unplanned conceptions who are conceived in times of better socioeconomic circumstances grow up in less stable families, although child later life outcomes are unaffected. The latter

finding is in contrast with the literature that finds enduring effects of family environment. However, it must be taken into account that this paper only focuses on the effect of parental marriage stability on the long-term outcomes more than fifty years later. Likewise, high equality of opportunity in the Netherlands could mitigate the influence of family environment on child long-term outcomes.

Thus, my findings do not directly provide a motive for policy aimed at reducing inequalities due to large changes in the socioeconomic environment at conception. However, this does not imply that policy-makers should not be cautious when it comes to these large shocks to the socioeconomic environment. There may be other potential drivers for parental selection, and future research should address how comparable shocks to the socioeconomic environment affect parental selection in different institutional settings and endeavor to identify effects earlier in the life-cycle.

Figures

Figure 2.1: The Netherlands by region, black circles indicate large cities in the west.

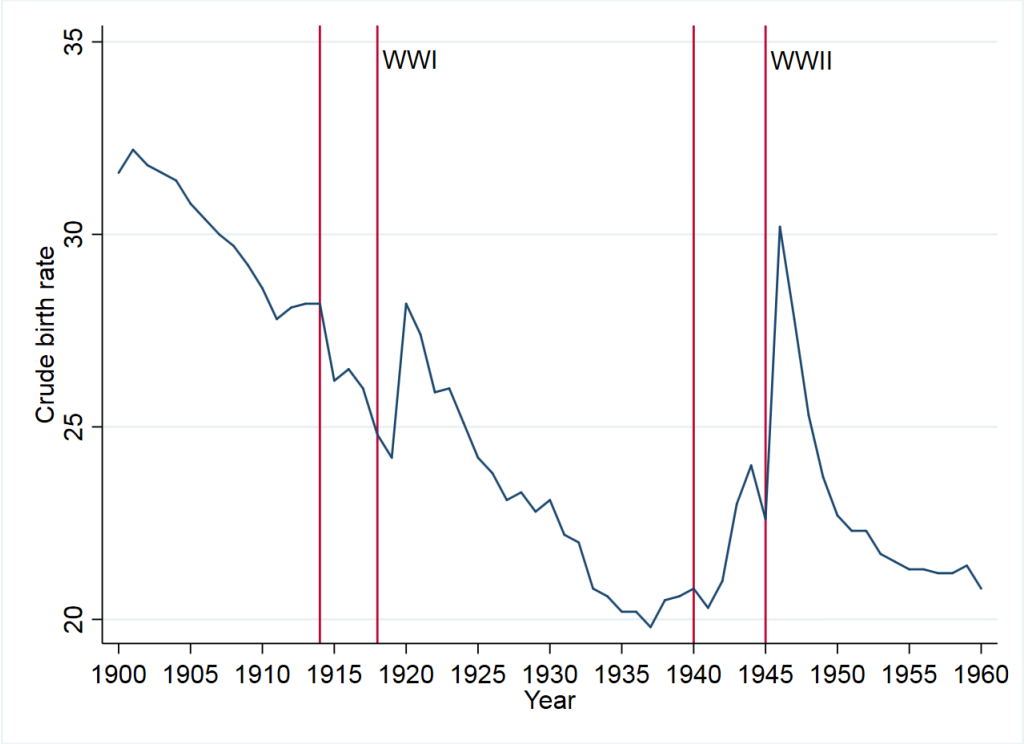


Figure 2.2: Yearly crude birth rate, number of live births per 1000 inhabitants, the Netherlands, 1900-1960. *Source:* Statistics Netherlands, statline.cbs.nl.

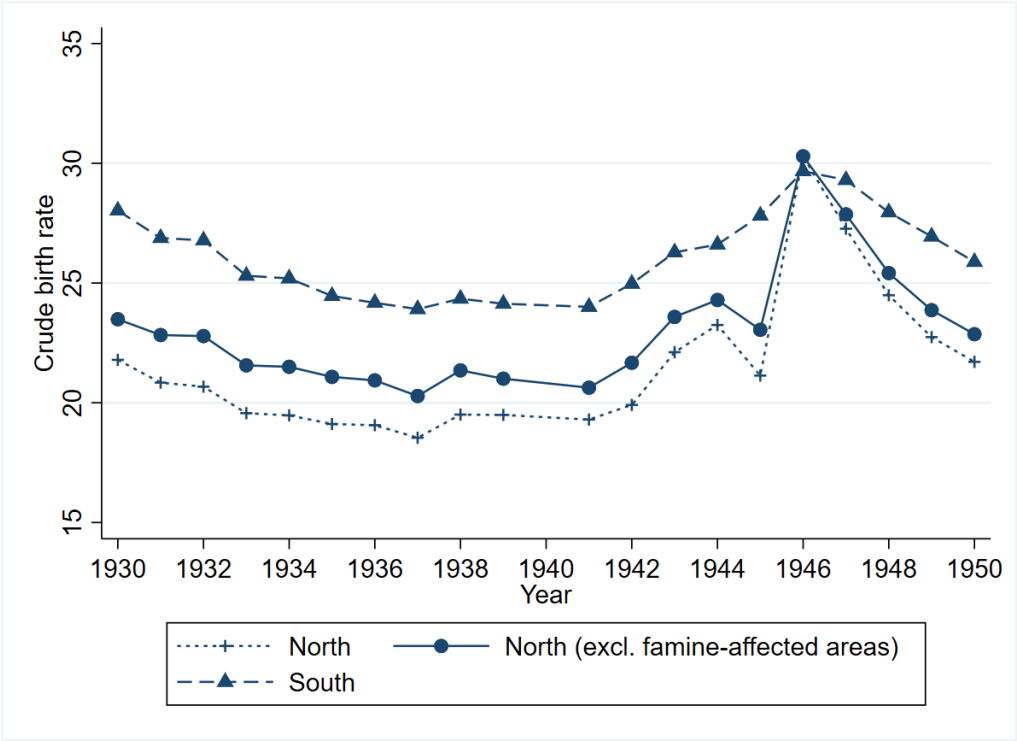


Figure 2.3: Yearly live births per 1000 inhabitants by region, the Netherlands, 1930-1950. North contains the provinces of Noord-Holland, Zuid-Holland, Utrecht, Groningen, Friesland, Drenthe, Overijssel, and Gelderland. South contains the provinces of Zeeland, Noord-Brabant, and Limburg. Data for 1940 is missing. *Source:* Author's calculations based on data from Statistics Netherlands Historical Collection, Loop van de bevolking per gemeente, historisch.cbs.nl.

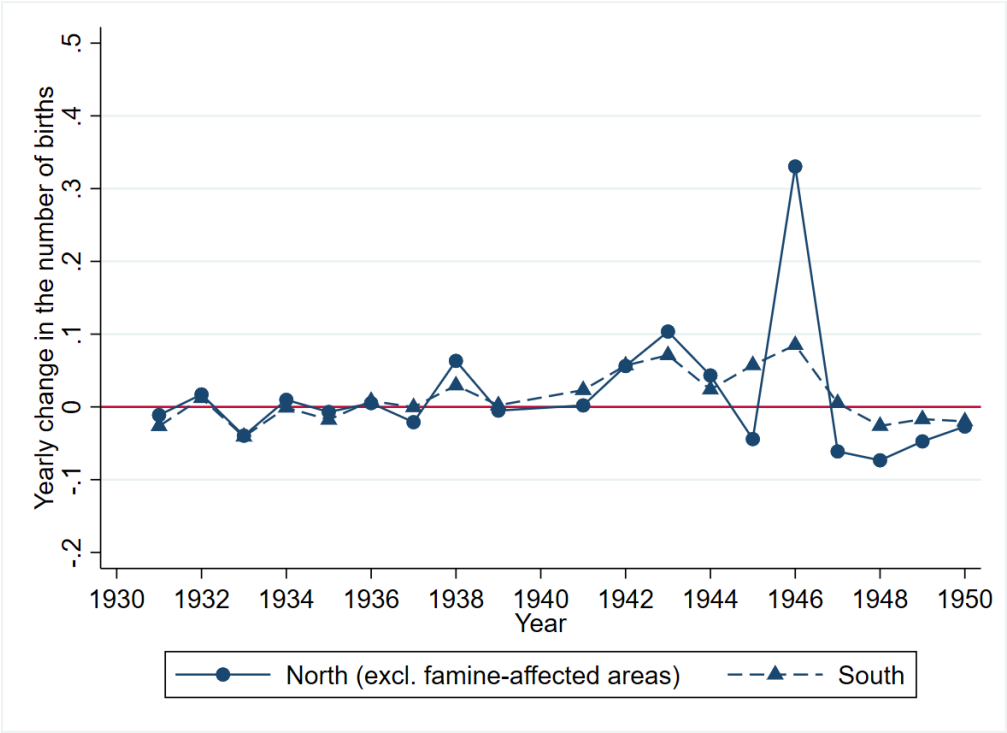


Figure 2.4: Yearly increase in the number of births by region, the Netherlands, 1931-1950. North contains the provinces of Noord-Holland, Zuid-Holland, Utrecht, Groningen, Friesland, Drenthe, Overijssel, and Gelderland. South contains the provinces of Zeeland, Noord-Brabant, and Limburg. Data for 1940 is missing. *Source:* Author's calculations based on data from Statistics Netherlands Historical Collection, Loop van de bevolking per gemeente, historisch.cbs.nl.

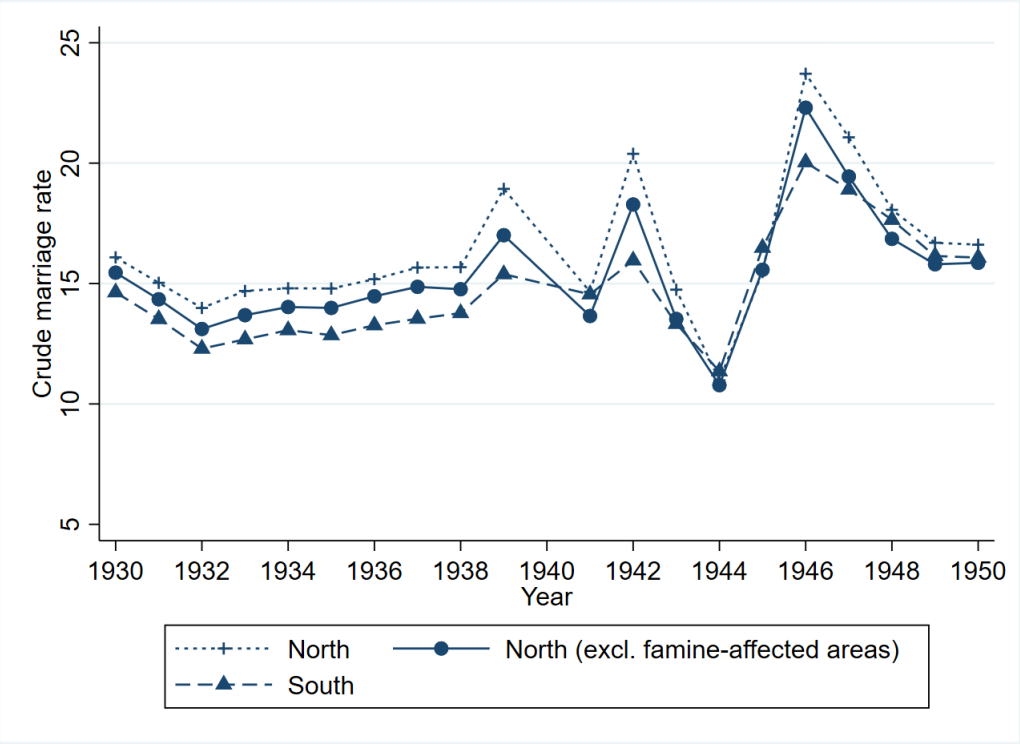


Figure 2.5: Yearly number of newly married individuals per 1000 inhabitants by region for 1930-1941, yearly number of marriages multiplied by two per 1000 inhabitants by region for 1942-1950, the Netherlands. Division of regions is the same as in Figure 2.3. *Source:* Author's calculations based on data from Statistics Netherlands Historical Collection, Loop van de bevolking per gemeente, historisch.cbs.nl.

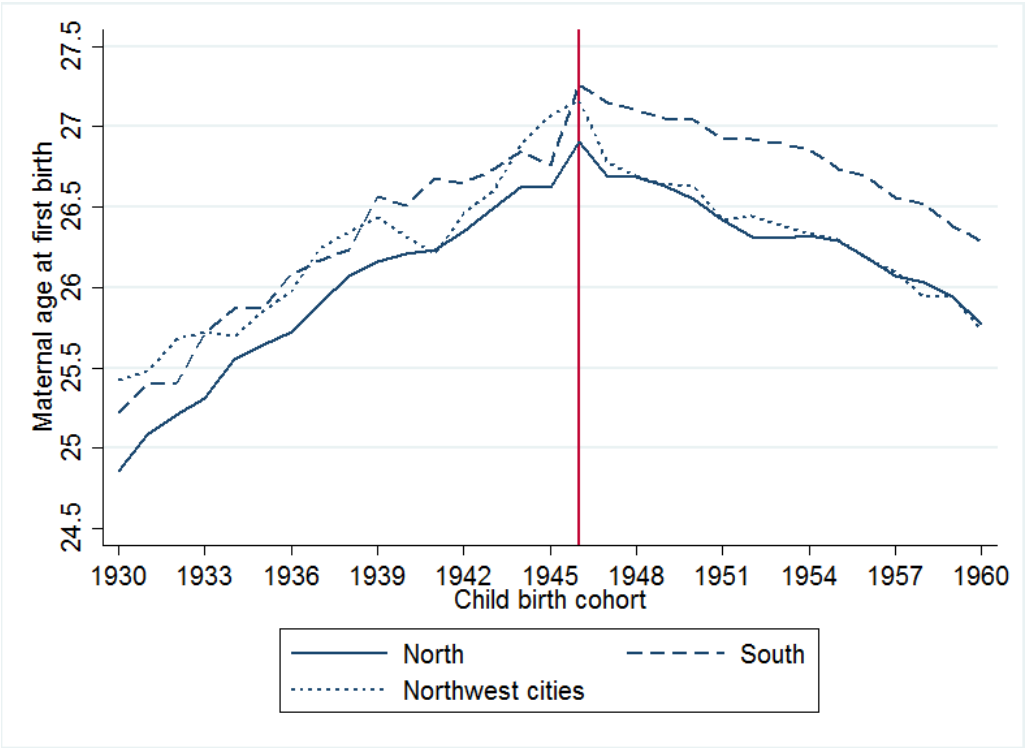


Figure 2.6: Yearly mean maternal age at first birth by child birth year and region, 1930-1960. North contains the northwest and northeast but excludes the large cities in the northwest of the Netherlands that suffered from the Hunger Winter. The trends for these cities are plotted separately. *Source:* Author's calculations based on administrative data from Statistics Netherlands, restricted sample.

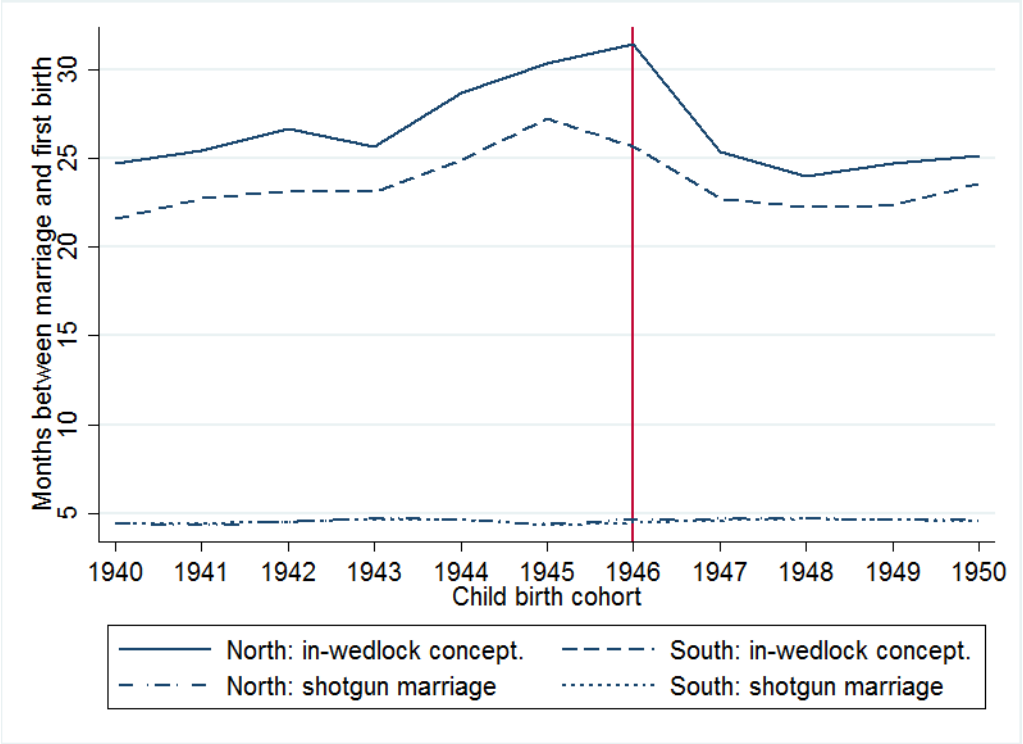


Figure 2.7: Yearly mean difference between marriage and first birth, in months, 1940-1950. North contains the northwest and northeast but excludes the large cities in the northwest of the Netherlands that suffered most from the Hunger Winter: Rotterdam, Amsterdam, The Hague, Delft, Utrecht and Haarlem. *Source:* Author's calculations based on administrative data from Statistics Netherlands, marital sample.

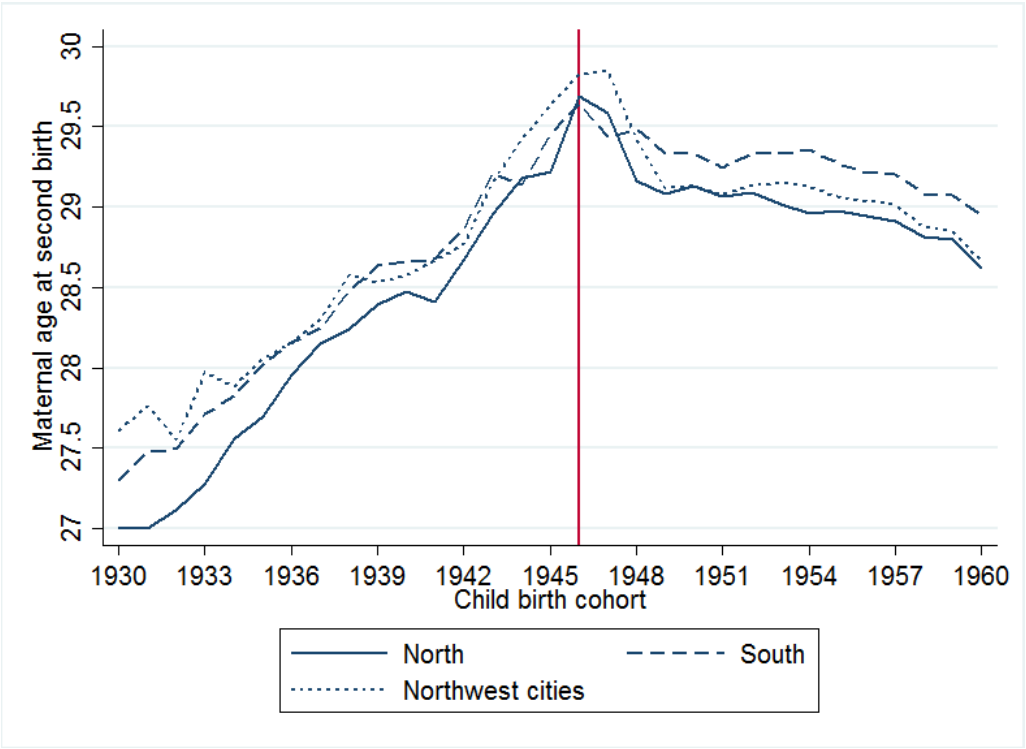


Figure 2.8: Yearly mean maternal age at second birth by child birth year and region, 1930-1960. North contains the northwest and northeast but excludes the large cities in the Northwest of the Netherlands that suffered most from the Hunger Winter. *Source:* Author's calculations based on administrative data from Statistics Netherlands, based on the restricted sample.

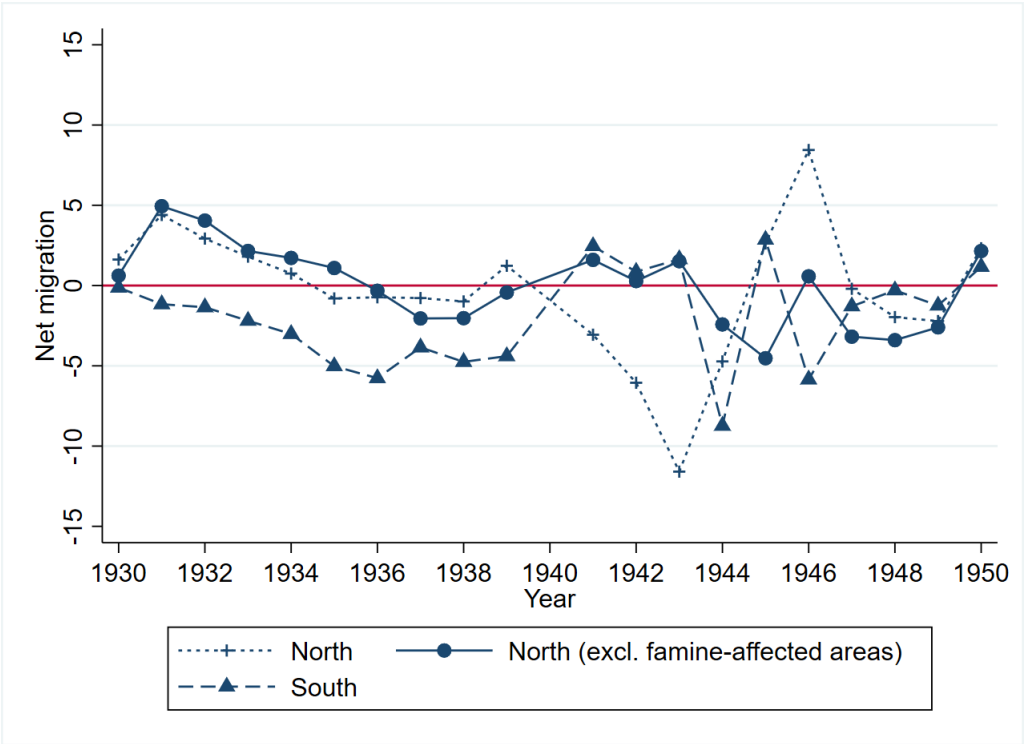


Figure 2.9: Net migration per 1000 inhabitants by region, the Netherlands, 1930-1950. Northwest contains the provinces of Noord-Holland, Zuid-Holland, and Utrecht. Northeast contains the provinces of Groningen, Friesland, Drenthe, Overijssel, and Gelderland. South contains the provinces of Zeeland, Noord-Brabant, and Limburg. *Source:* Author's calculations based on data from Statistics Netherlands Historical Collection, historisch.cbs.nl.

Figure 2.10: GDP total production in million guilders represented on the left axis. Industrial employment in thousands of employees represented on the right axis, The Netherlands, 1938-1948. *Source:* Klemann (2002)

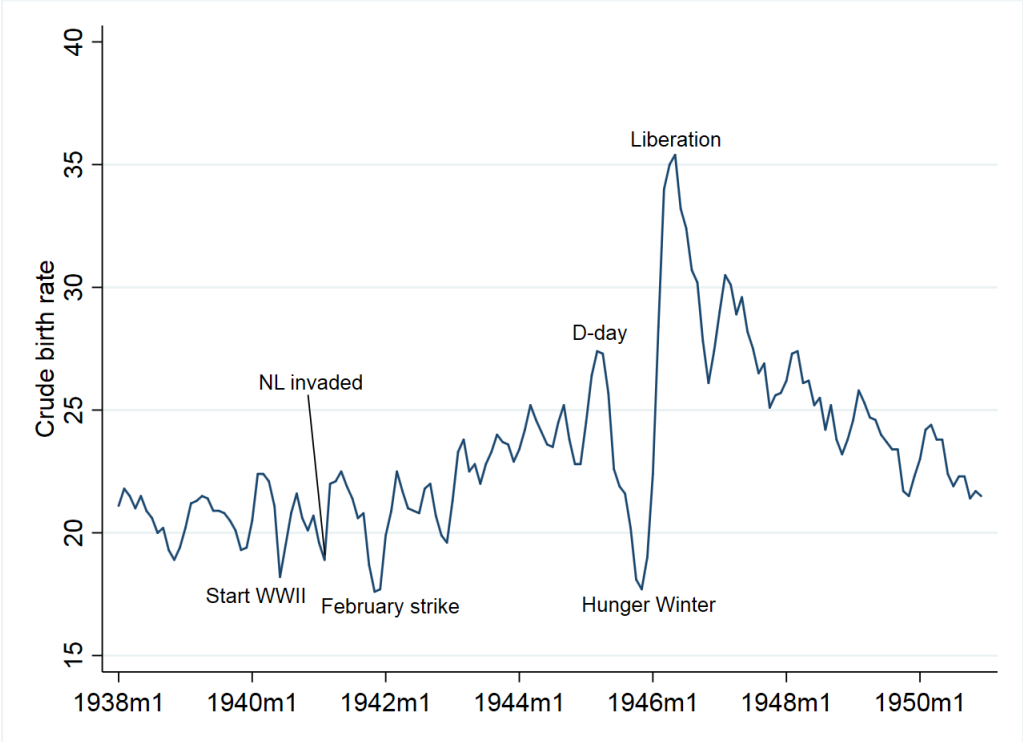


Figure 2.11: Monthly crude birth rate, the Netherlands, 1938-1950. *Source:* Statistics Netherlands Historical Collection, historisch.cbs.nl, Statistiek van de loop van de bevolking 1938, p. 45.

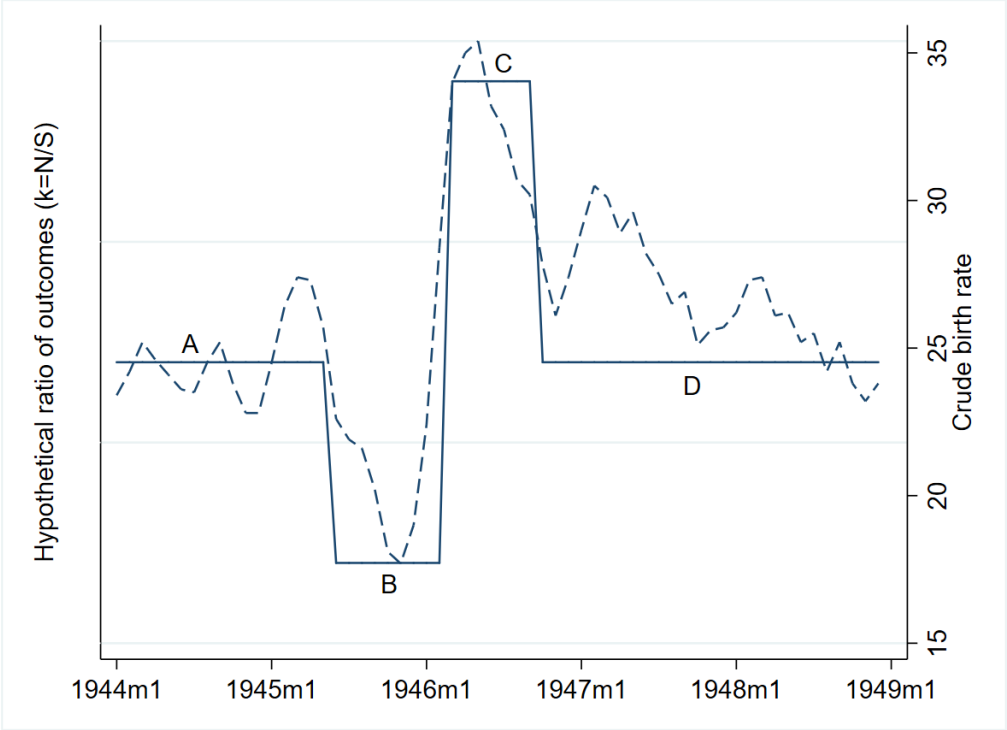


Figure 2.12: Hypothetical ratio (k) of child outcomes in north vs south (north/south) depicted on left axis. Monthly crude birth rate depicted on right axis.

Tables

Table 2.1: Frequencies by cohort and birth region

	Unrestricted sample		Restricted sample		Marital sample	
	Males	Females	Males	Females	Males	Females
Pre-Birth Peak cohort						
North	104,993	100,695	67,961	64,816	57,446	55,738
South	50,568	48,657	30,089	28,955	25,177	24,646
NW cities	39,259	37,976	26,646	26,293	21,974	21,679
Birth Peak cohort						
North	40,866	39,248	28,114	26,784	24,079	23,377
South	16,590	15,887	10,534	10,127	8,950	8,750
NW cities	21,861	20,778	15,916	15,325	13,334	12,764
Post-Birth Peak cohort						
North	132,015	126,712	95,292	90,831	83,224	80,398
South	60,855	58,222	41,414	39,455	35,965	34,701
NW cities	63,106	60,425	47,057	45,536	40,091	38,770
Totals						
All	530,113	508,600	363,023	348,122	310,240	300,823
Without NW cities	405,887	389,421	273,404	260,968	234,841	227,610

Notes: Frequencies observed in the sample as described in Section 4.4, by birth region and sample. Children in the Birth Peak cohort are born between March and September 1946. The pre-Birth Peak cohort contains individuals born from January 1944 up until the Birth Peak cohort. The post-Birth Peak cohort contains individuals born after the BP cohort until 1948. The right column shows the number of children observed to be born in-wedlock by birth year and birth region (within the restricted sample).

Table 2.2: Descriptive statistics

	North (1)	South (2)	Δ (1) - (2)
# individuals	187,675	79,705	
Birth cohort			
Pre-Birth Peak	132,777	59,044	
1946 (March-September)	54,898	20,661	
By parental marital status			
In marital sample	160,640	67,523	
First birth: In-wedlock conception	45,455	19,630	
First birth: Shotgun marriage	11,493	2,897	
Family characteristics			
Maternal age at birth	29.445	29.874	***
Paternal age at birth	32.392	32.770	***
Urban	0.108	0.248	***
Birth order	2.281	2.439	***
Siblings	4.271	4.704	***
Length parental marriage (days)	16,828.93	16,566.30	***
Health outcomes			
Mortality ≤ 65	0.083	0.084	
	(0.277)	(0.278)	
Mortality ≤ 70	0.134	0.135	
	(0.341)	(0.342)	
Any drugs (0/1)	0.477	0.485	***
	(0.499)	(0.500)	
Labor market outcomes			
# individuals (males only)	96,075	40,623	
Employment (0/1)	0.841	0.819	***
	(0.366)	(0.385)	
Labor income (ln)*	7.738	7.423	***
	(5.651)	(5.889)	
Enrollment in social security (0/1)	0.230	0.252	***
	(0.421)	(0.434)	

Notes: Descriptive statistics for the restricted sample, see data-section. Standard deviations between brackets.

Table 2.3: Descriptive statistics: first births by marital state at conception

	In-wedlock conceptions (1)	Shotgun conceptions (2)	Δ (1) - (2)
# individuals	65,085	14,390	
Maternal age at birth	27.204	22.821	***
Paternal age at birth	30.077	24.853	***
Urban	0.166	0.116	***
Birth order	1	1	
Siblings	3.348	3.730	***
Length parental marriage (days)	16,455.42	16,571.70	**
Health outcomes			
Mortality ≤ 65	0.081	0.085	*
	(0.272)	(0.279)	
Mortality ≤ 70	0.129	0.139	***
	(0.335)	(0.346)	
Any drugs (0/1)	0.475	0.502	***
	(0.499)	(0.500)	
Labor market outcomes			
# individuals (males only)	33,186	7,442	
Employment (0/1)	0.844	0.829	***
	(0.363)	(0.376)	
Labor income (ln)*	7.853	7.567	***
	(5.635)	(5.751)	
Enrollment in social security (0/1)	0.201	0.238	***
	(0.401)	(0.426)	

Notes: Descriptive statistics for the restricted sample, focus on first born children who were born in-wedlock, distinction between in-wedlock conceptions and shotgun conceptions (see data-section for more information). Standard deviations between brackets. The last column presents the outcomes of a t-test on the equality of means.

Table 2.4: Family characteristics

	Family size		Marriage stability	
	(1)	(2)	(3)	(4)
Panel A: Pre-sample				
BP	-0.092***	-0.001	1.88	57.52
	(0.028)	(0.052)	(39.99)	(76.07)
North	-0.395***	-0.432***	247.04***	168.23**
	(0.022)	(0.040)	(31.70)	(69.23)
CoBP	-0.117***	-0.274***	52.36	54.68
	(0.042)	(0.072)	(52.10)	(64.65)
Controls	No	Yes	No	Yes
Mean dep. var.	4.400	4.400	16,751.34	16,751.34
N	267,380	267,380	221,999	221,999
Panel B: By marital status at conception				
BP	-0.183	-0.028	85.08	385.57
	(0.115)	(0.260)	(219.50)	(500.67)
North	-0.317***	-0.157	267.68	634.79
	(0.057)	(0.231)	(183.38)	(385.94)
CoBP	-0.311*	-0.037	88.57	-935.31**
	(0.158)	(0.224)	(168.32)	(435.85)
Controls	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG
Mean dep. var.	3.348	3.730	16,455.42	16,571.70
N	65,085	14,390	65,085	14,390

Notes: Estimated by OLS. Standard errors are clustered by birth month/year and birth region (north versus south). The specification contains region-specific birth-month/year trends (linear and quadratic), an indicator for being conceived during the liberation of the south, birth month dummies, and a control for maternal age at birth.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.5: Labor market outcomes - Males

	Employment		Labor earnings (ln)		Social security	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pre-sample						
BP	0.026***	0.005	0.403***	0.029	-0.037***	0.002
	(0.004)	(0.010)	(0.060)	(0.165)	(0.005)	(0.013)
North	0.022***	0.032***	0.314***	0.459***	-0.025***	-0.039***
	(0.004)	(0.007)	(0.065)	(0.115)	(0.005)	(0.006)
CoBP	-0.004	-0.007	-0.044	-0.117	0.014**	0.003
	(0.005)	(0.009)	(0.079)	(0.129)	(0.007)	(0.012)
Controls	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.834	0.834	7.644	7.644	0.237	0.237
N	136,698	136,698	136,698	136,698	136,698	136,698
Panel B: Pre-sample by marital status at conception						
BP	-0.031	0.037	-0.466	0.211	0.015	0.045
	(0.023)	(0.038)	(0.356)	(0.579)	(0.020)	(0.064)
North	0.028***	0.065*	0.426**	0.801	-0.026**	-0.082*
	(0.010)	(0.033)	(0.165)	(0.527)	(0.011)	(0.045)
CoBP	-0.006	-0.009	-0.201	0.169	-0.006	-0.058
	(0.021)	(0.032)	(0.328)	(0.488)	(0.018)	(0.052)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.844	0.829	7.853	7.567	0.201	0.238
N	33,186	7,442	33,186	7,442	33,186	7,442

Notes: Estimated by OLS, males only, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and controls for whether the individuals was conceived during the liberation of the south, and whether the individual is born in a city, and maternal age at birth. IW indicates that the individual is conceived in-wedlock, whereas SG denotes a shotgun conception.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.6: Health outcomes

	Mortality ≤ 65		Mortality ≤ 70		I(Any drugs)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pre-sample						
BP	0.002	0.001	-0.000	0.003	-0.014**	0.035***
	(0.002)	(0.005)	(0.003)	(0.006)	(0.005)	(0.006)
North	-0.000	0.000	-0.000	0.000	-0.006	-0.003
	(0.001)	(0.005)	(0.002)	(0.005)	(0.005)	(0.007)
CoBP	-0.003	-0.002	-0.004	-0.005	-0.006	-0.024***
	(0.002)	(0.005)	(0.003)	(0.005)	(0.006)	(0.007)
Controls	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.084	0.084	0.135	0.135	0.479	0.479
N	267,380	267,380	267,380	267,380	267,380	267,380
Panel B: Pre-sample by marital status at conception						
BP	-0.011	-0.019	-0.013	0.012	0.044**	0.059
	(0.011)	(0.032)	(0.011)	(0.043)	(0.016)	(0.045)
North	-0.005	0.008	-0.006	-0.016	-0.005	-0.037
	(0.008)	(0.017)	(0.008)	(0.029)	(0.014)	(0.023)
CoBP	0.011	0.032	0.009	0.009	-0.049***	-0.070*
	(0.010)	(0.027)	(0.011)	(0.036)	(0.016)	(0.041)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.081	0.085	0.129	0.139	0.475	0.502
N	65,085	14,390	65,085	14,390	65,085	14,390

Notes: Estimated by OLS, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and controls for whether the individuals is born in a city, and maternal age at birth. The pre-sample specification also contains an indicator for being conceived during the liberation of the south. IW indicates that the individual is conceived in-wedlock, whereas SG denotes a shotgun conception.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.7: Robustness results - Family outcomes

	Family size	Length marriage
	(1)	(2)
Baseline	-0.274***	54.68
	(0.072)	(64.65)
N	267,380	221,999
Post-cohorts	-0.036	-24.78
	(0.055)	(112.83)
N	342,551	294,189
First birth only	-0.242**	-71.372
	(0.120)	(128.038)
N	99,855	79,475
Closer to the rivers	-0.338***	165.493*
	(0.066)	(88.243)
N	167,155	138,710
Treatment: Incl. Feb	-0.302***	61.95
	(0.064)	(61.96)
N	267,380	221,999
Treatment: Mar-May	-0.272***	101.50
	(0.063)	(67.32)
N	225,770	186,833
Donut: Hunger Winter	-0.162***	-4.78
	(0.039)	(50.33)
N	186,402	155,546
Donut: south liberated	-0.377***	-39.46
	(0.089)	(152.90)
N	202,120	168,407

Notes : Double-difference estimate Birth'46*North is reported, full control specification.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.8: Robustness results - Child outcomes

	Employ.	Labor earn. (ln)	Social Security	Mort.≤ 65	Mort. ≤ 70	I(Any drugs)
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline	-0.007	-0.117	0.003	-0.002	-0.005	-0.024***
	(0.009)	(0.129)	(0.012)	(0.005)	(0.005)	(0.007)
N	136,698	136,698	136,698	267,380	267,380	267,380
Post-cohorts	-0.012	-0.164	0.014*	0.002	-0.003	-0.005
	(0.008)	(0.130)	(0.008)	(0.003)	(0.005)	(0.007)
N	175,354	175,354	175,354	342,551	342,551	342,551
First birth only	0.005	0.035	-0.016	0.010	0.003	-0.044***
	(0.019)	(0.284)	(0.015)	(0.008)	(0.010)	(0.014)
N	51,272	51,272	51,272	99,855	99,855	99,855
Males only				-0.006	-0.006	-0.020
				(0.007)	(0.009)	(0.012)
N				136,698	136,698	136,698
Females only				0.001	-0.005	-0.028**
				(0.006)	(0.007)	(0.012)
N				130,682	130,682	130,682
Out-wedlock birth	0.042	0.805	0.025	-0.025	-0.045	0.019
	(0.045)	(0.721)	(0.081)	(0.027)	(0.034)	(0.043)
N	3,063	3,063	3,063	6,164	6,164	6,164
Closer to the rivers	-0.010	-0.192	0.002	-0.008	-0.009*	-0.016*
	(0.008)	(0.125)	(0.013)	(0.005)	(0.005)	(0.009)
N	85,389	85,389	85,389	167,155	167,155	167,155
Treatment: Incl. Feb	-0.014*	-0.200	0.004	0.002	-0.006	-0.008
	(0.008)	(0.126)	(0.010)	(0.005)	(0.005)	(0.010)
N	136,698	136,698	136,698	267,380	267,380	267,380
Treatment: Mar-May	-0.005	-0.077	-0.003	-0.003	-0.003	-0.025***
	(0.009)	(0.135)	(0.012)	(0.005)	(0.005)	(0.008)
N	115,353	115,353	115,353	225,770	225,770	225,770
Donut: Hunger Winter	-0.008	-0.165	0.006	-0.003	-0.008	-0.018*
	(0.008)	(0.123)	(0.011)	(0.005)	(0.005)	(0.008)
N	95,350	95,350	95,350	186,402	186,402	186,402
Donut: south liberated	-0.015	-0.283	0.030	0.005	-0.015	-0.003
	(0.011)	(0.189)	(0.021)	(0.010)	(0.012)	(0.010)
N	103,331	103,331	103,331	202,120	202,120	202,120
Conditioning on mortality	-0.006	-0.102	0.003	-0.010**	-0.013**	-0.021***
	(0.008)	(0.118)	(0.012)	(0.004)	(0.005)	(0.007)
N	134,500	134,500	134,500	253,837	253,837	253,837

Notes : Double-difference estimate Birth'46*North is reported, full control specification.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.9: CoBP siblings - males - labor market outcomes

	Employment		Labor earnings (ln)		Social security	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:						
Sibling in BP	-0.010***	-0.014**	-0.146***	-0.197**	0.014***	0.011
	(0.003)	(0.006)	(0.052)	(0.093)	(0.005)	(0.007)
North	0.031***	0.033**	0.587***	0.635***	-0.046***	-0.044**
	(0.011)	(0.013)	(0.179)	(0.206)	(0.016)	(0.019)
CoBP Siblings	0.004	0.011	0.079	0.166	-0.003	-0.005
	(0.004)	(0.007)	(0.066)	(0.106)	(0.006)	(0.008)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
2-child families	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.880	0.878	8.341	8.330	0.190	0.188
N	139,247	90,599	139,247	90,599	139,247	90,599
Panel B: By marital status at first conception						
Sibling in BP	-0.015**	0.002	-0.232**	0.120	0.012	0.009
	(0.007)	(0.017)	(0.103)	(0.252)	(0.008)	(0.021)
North	0.010	0.080**	0.273	1.126*	-0.010	-0.076
	(0.014)	(0.036)	(0.234)	(0.578)	(0.022)	(0.046)
CoBP Siblings	0.008	0.019	0.147	0.231	-0.002	-0.025
	(0.008)	(0.018)	(0.119)	(0.269)	(0.009)	(0.023)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
2-child families	Yes	Yes	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.889	0.858	8.528	7.965	0.168	0.231
N	63,690	14,686	63,690	14,686	63,690	14,686

Notes: Estimated by OLS, males only, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, whether the individuals is born in a city, and maternal age at birth.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.10: CoBP siblings - health outcomes

	Mortality ≤ 65		Mortality ≤ 70		I(Any drugs)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:						
Sibling in BP	-0.003	0.001	-0.004	0.001	0.008**	0.017***
	(0.002)	(0.004)	(0.003)	(0.004)	(0.004)	(0.005)
North	0.001	0.007	0.009	0.012	-0.006	-0.002
	(0.007)	(0.009)	(0.009)	(0.011)	(0.013)	(0.015)
CoBP Siblings	0.005*	0.003	0.007**	0.005	-0.002	-0.003
	(0.003)	(0.004)	(0.003)	(0.005)	(0.005)	(0.007)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
2-child families only	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.084	0.085	0.119	0.122	0.419	0.427
N	272,644	176,834	272,644	176,834	272,644	176,834
Panel B: By marital status at first conception						
Sibling in BP	0.005	-0.015	0.004	-0.020	0.013**	0.005
	(0.004)	(0.010)	(0.005)	(0.012)	(0.006)	(0.017)
North	0.016*	-0.022	0.023*	-0.033	-0.008	-0.031
	(0.009)	(0.027)	(0.012)	(0.032)	(0.017)	(0.043)
CoBP Siblings	-0.003	0.020*	0.001	0.018	0.012	-0.027
	(0.005)	(0.011)	(0.006)	(0.014)	(0.008)	(0.019)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
2-child families	No	Yes	No	Yes	No	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.080	0.093	0.114	0.132	0.415	0.456
N	124,798	28,877	124,798	28,877	124,798	28,877

Notes: Estimated by OLS, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, whether the individuals is born in a city, and maternal age at birth.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.11: Selection in restricted sample

	I(Mother observed in restricted sample)			
	Pre-sample		Post-sample	
	(1)	(2)	(3)	(4)
BP	0.041***	-0.000	-0.043***	-0.008
	(0.005)	(0.009)	(0.005)	(0.006)
North	0.050***	0.035***	0.040***	0.025**
	(0.005)	(0.005)	(0.005)	(0.010)
CoBP	-0.001	0.003	0.009	0.007
	(0.007)	(0.009)	(0.007)	(0.007)
Mat. age at birth		-0.027***		-0.023***
		(0.000)		(0.000)
Controls	No	Yes	No	Yes
Mean value of outcome	0.640	0.640	0.699	0.699
N	417,504	417,504	490,395	490,395

Notes: Estimated by OLS. Standard errors are clustered by birth month/year and birth region (north versus south). The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies and a control for maternal age at birth. The pre-sample specification also contains an indicator for being conceived during the liberation of the south. Estimated for unrestricted sample.

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

Table 2.12: Mechanisms at work

	Employ.	Labor earn. (ln)	Social Security	Mort.≤ 65	Mort. ≤ 70	I(Any drugs)
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline	-0.007	-0.117	0.003	-0.002	-0.005	-0.024***
	(0.009)	(0.129)	(0.012)	(0.005)	(0.005)	(0.007)
N	136,698	136,698	136,698	267,380	267,380	267,380
Cohort effects or parental selection?						
Cohort size	-0.010	-0.177	0.010	-0.002	-0.004	-0.022***
	(0.009)	(0.135)	(0.012)	(0.005)	(0.005)	(0.007)
N	136,698	136,698	136,698	267,380	267,380	267,380
Influence of the Hunger Winter						
Incl. HW-cities	-0.005	-0.057	-0.000	-0.003	-0.006	-0.022***
	(0.009)	(0.137)	(0.012)	(0.004)	(0.004)	(0.007)
N	179,260	179,260	179,260	351,560	351,560	351,560
Influence of forced labor absence						
Excl. border prov.	-0.014	-0.205	0.009	-0.001	-0.005	-0.021**
	(0.011)	(0.142)	(0.011)	(0.007)	(0.008)	(0.009)
N	76,443	76,443	76,443	149,508	149,508	149,508
Influence of Canadian Liberators						
Leave-dates	0.004	0.117	-0.020	-0.007	0.011	-0.034**
	(0.016)	(0.244)	(0.016)	(0.007)	(0.008)	(0.013)
N	108,441	108,441	108,441	211,998	211,998	211,998
Leave-centres	-0.011	-0.223	0.010	-0.002	-0.009	-0.023***
	(0.009)	(0.145)	(0.014)	(0.004)	(0.006)	(0.008)
N	106,896	106,896	106,896	209,098	209,098	209,098
Sample selection effects						
Unrestricted sample	-0.006	-0.054	-0.005	-0.004	-0.008	-0.016**
	(0.009)	(0.136)	(0.010)	(0.004)	(0.005)	(0.007)
N	213,017	213,017	213,017	417,504	417,504	417,504

Notes : Double-difference estimate Birth'46*North is reported, full control specification.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix A: Additional Figures

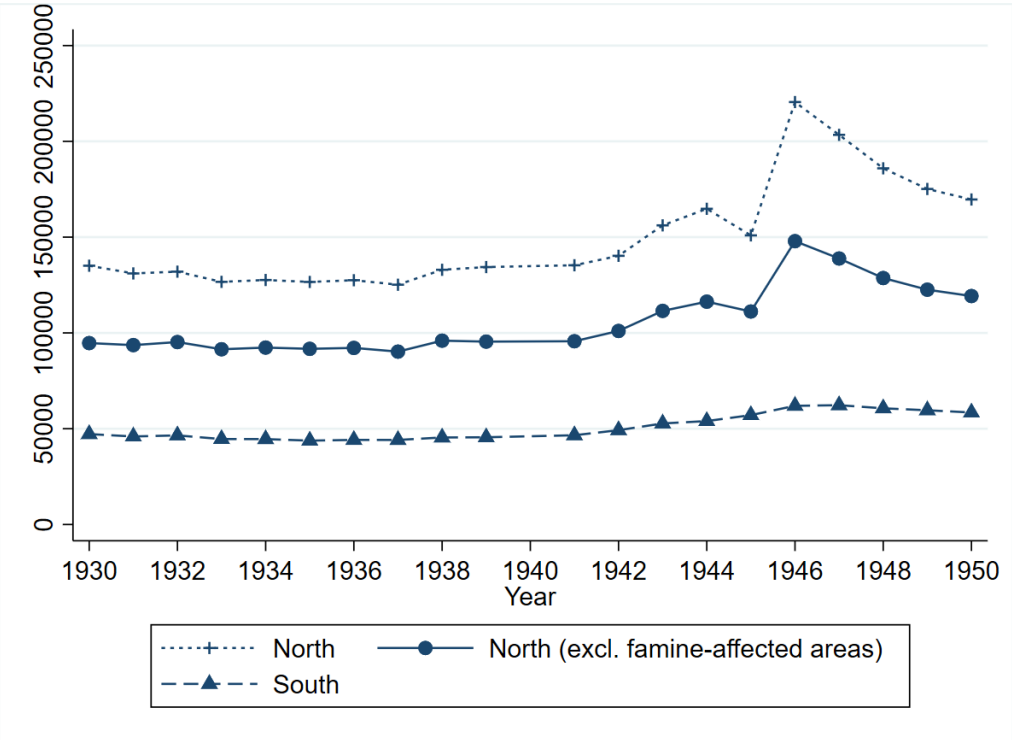


Figure 2.A1: Yearly live births by region, the Netherlands, 1930-1950. North contains the provinces of Noord-Holland, Zuid-Holland, Utrecht, Groningen, Friesland, Drenthe, Overijssel, and Gelderland. South contains the provinces of Zeeland, Noord-Brabant, and Limburg. Data for 1940 is missing. *Source:* Author's calculations based on data from Statistics Netherlands Historical Collection, Loop van de bevolking per gemeente, historisch.cbs.nl.

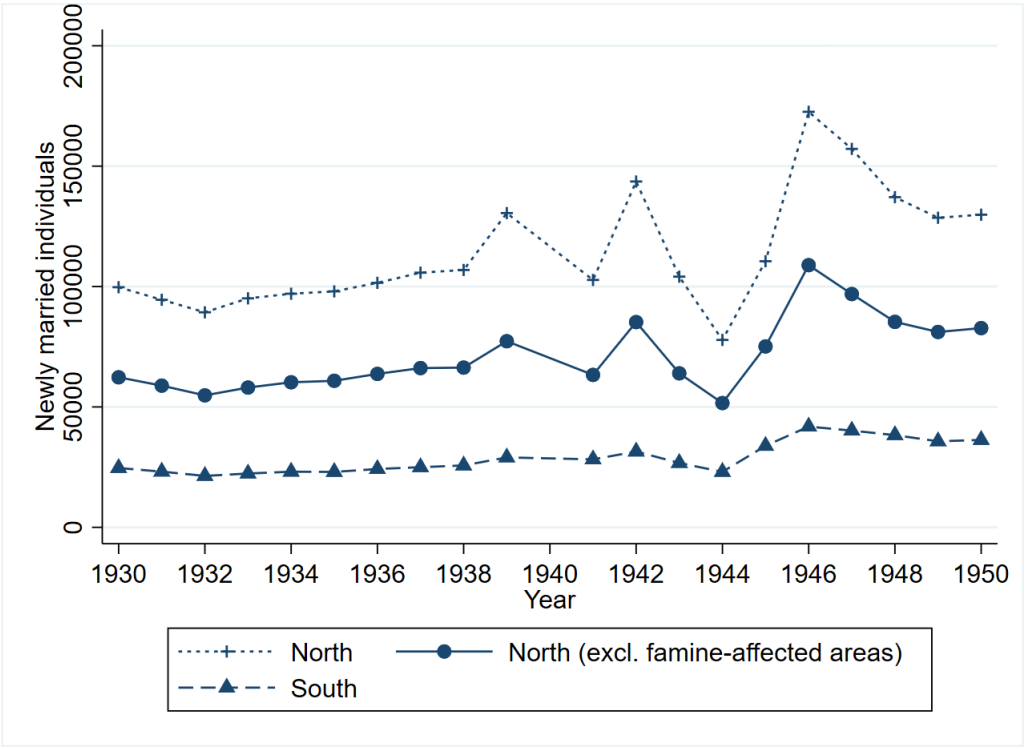


Figure 2.A2: Yearly number of newly married individuals by region for 1930-1941, yearly number of marriages multiplied by two by region for 1942-1950, the Netherlands. Division of regions is the same as in Figure 2.3. *Source:* Author's calculations based on data from Statistics Netherlands Historical Collection, Loop van de bevolking per gemeente, historisch.cbs.nl.

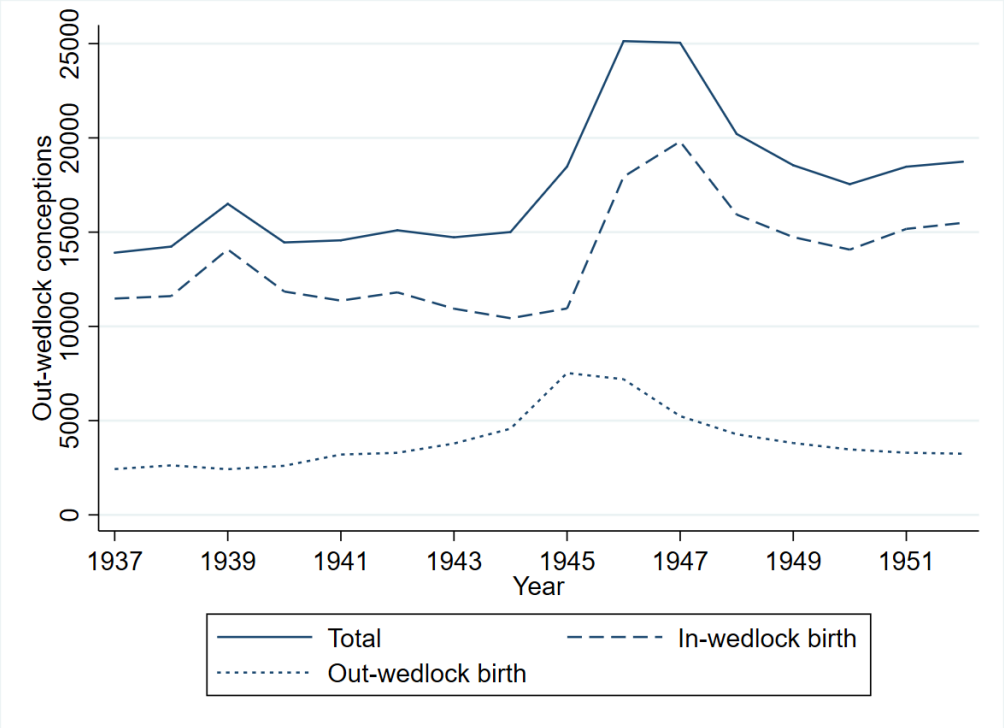


Figure 2.A3: Yearly number of out-wedlock conceptions, separated by whether they are followed by a marriage (in-wedlock birth) or not (out-wedlock birth). *Source:* CBS (1975).

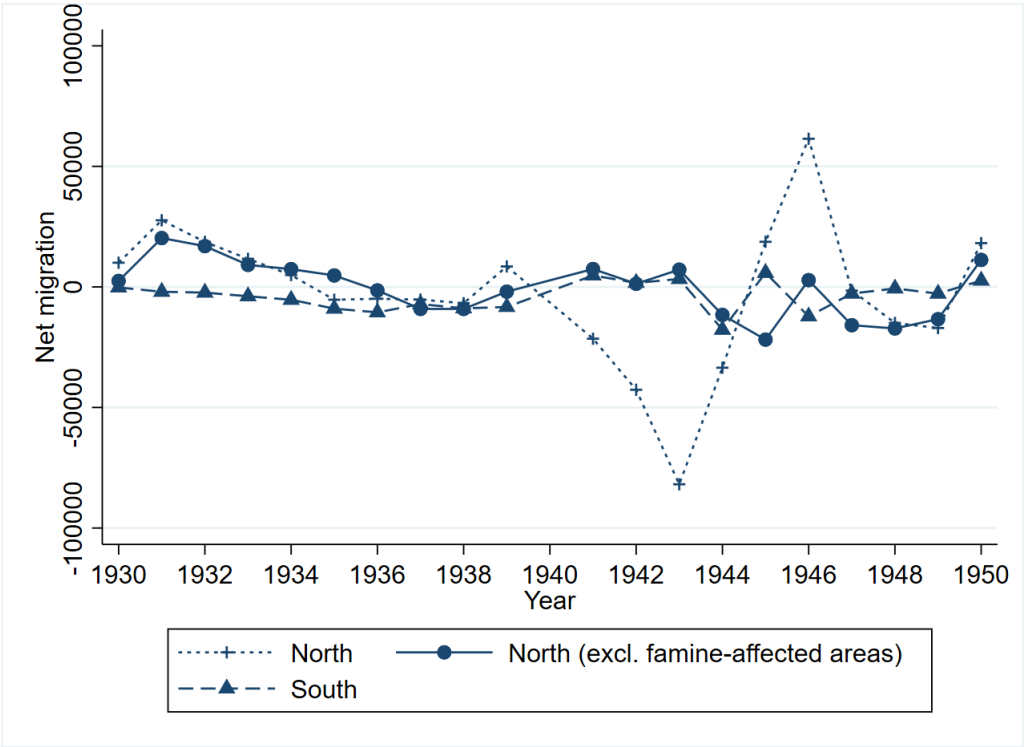


Figure 2.A4: Net migration by region, the Netherlands, 1930-1950. Division of regions is the same as in Figure 2.3. *Source:* Author's calculations based on data from Statistics Netherlands Historical Collection, historisch.cbs.nl.

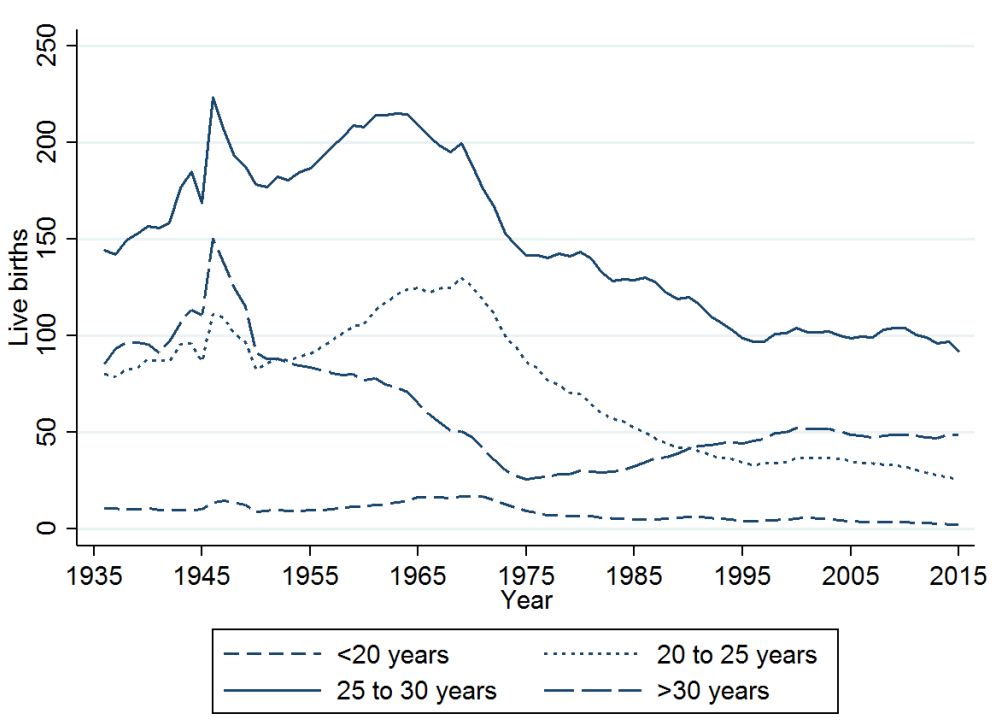


Figure 2.A5: Yearly live births per 1000 inhabitants by maternal age at birth, the Netherlands, 1936-2015. *Source:* Statistics Netherlands, statline.cbs.nl.

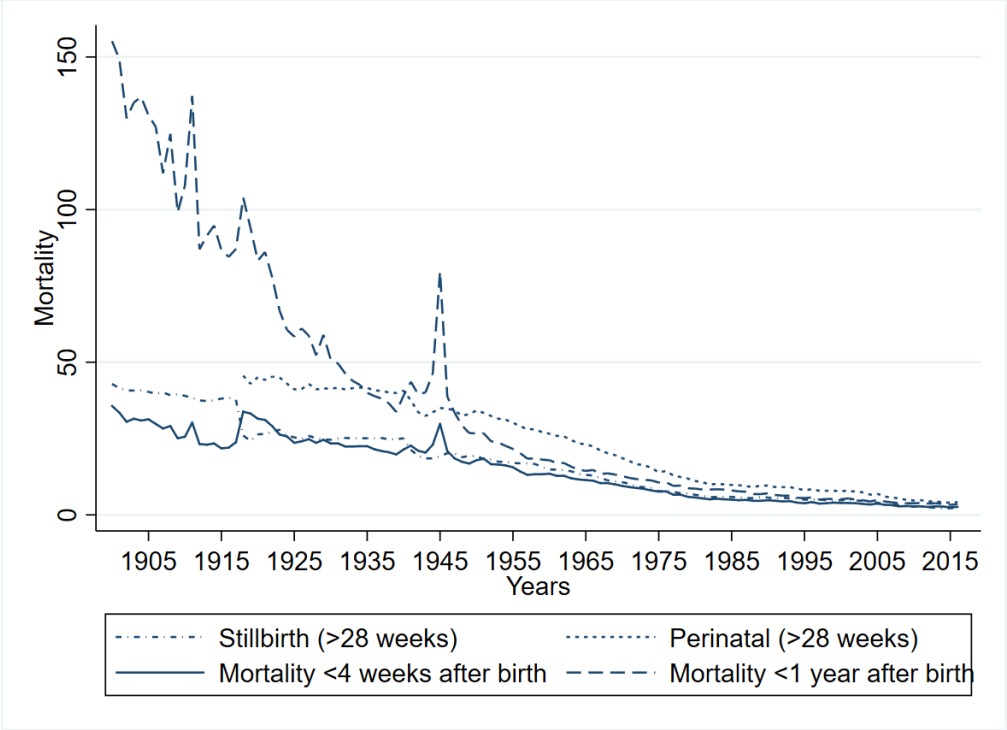


Figure 2.A6: Yearly mortality per 1000 live births for perinatal deaths and deaths after birth, yearly mortality per 1000 births for stillbirths, the Netherlands, 1900-2016. *Source:* Statistics Netherlands, statline.cbs.nl.

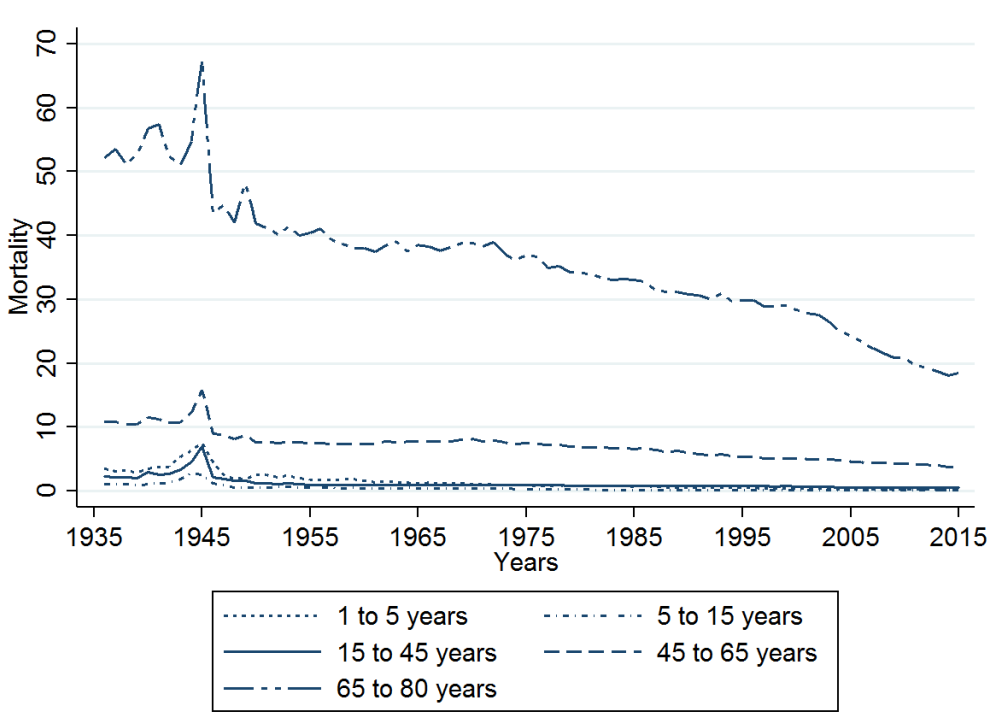


Figure 2.A7: Yearly mortality per 1000 inhabitants by age, the Netherlands, 1936-2015. *Source:* Statistics Netherlands, statline.cbs.nl.

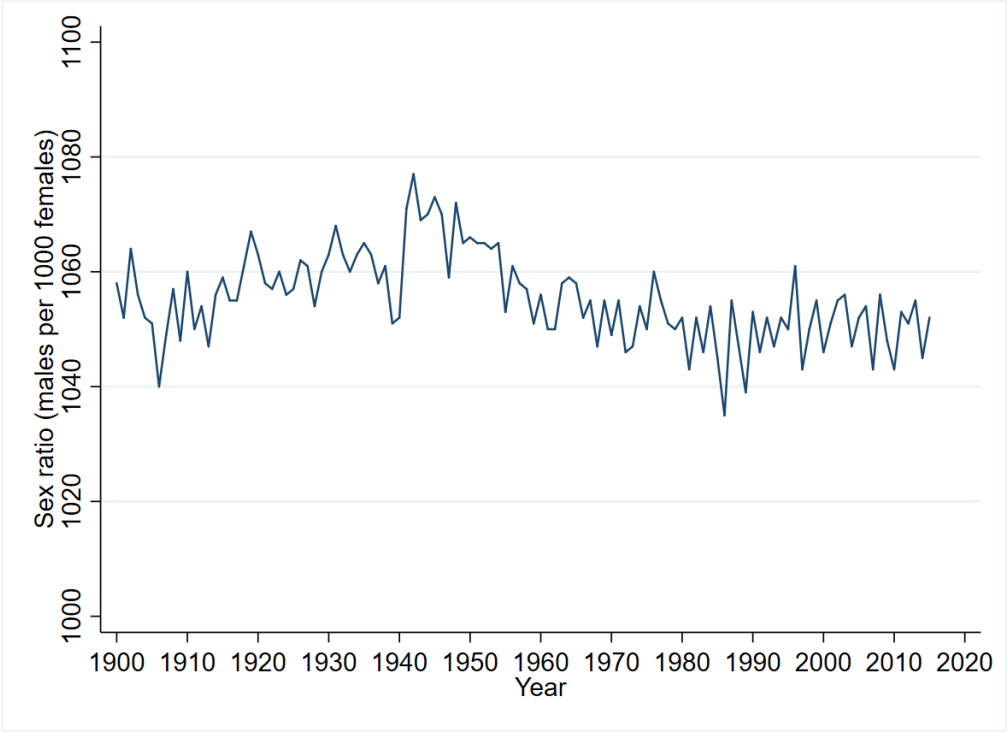
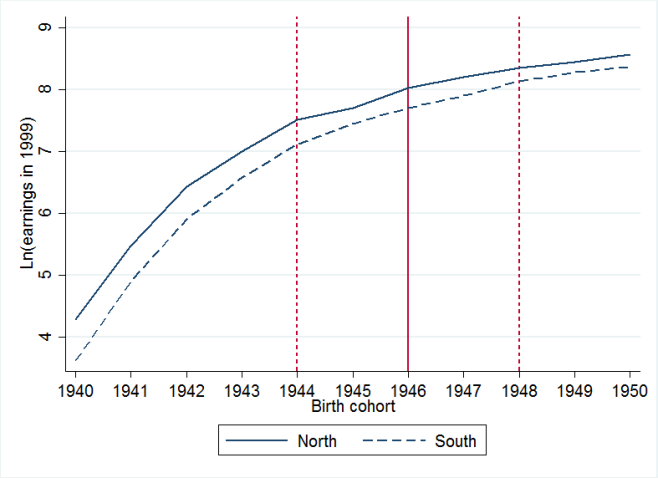
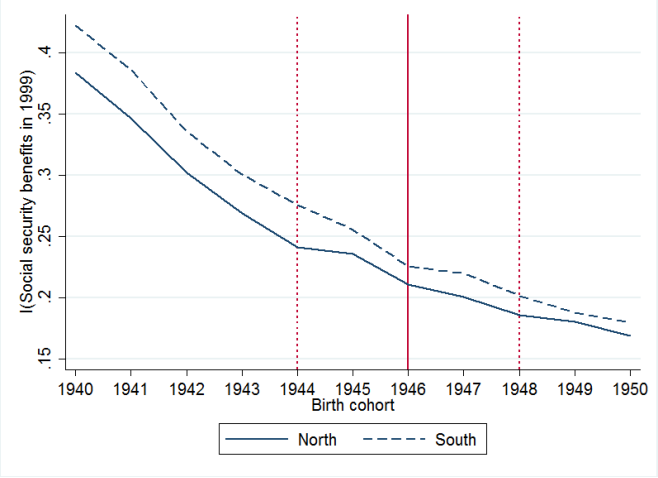


Figure 2.A8: Sex ratio, yearly number of males born for every 1000 females, the Netherlands, 1900-2015. *Source:* Statistics Netherlands, statline.cbs.nl.

(a) Employment in 1999 (0/1)

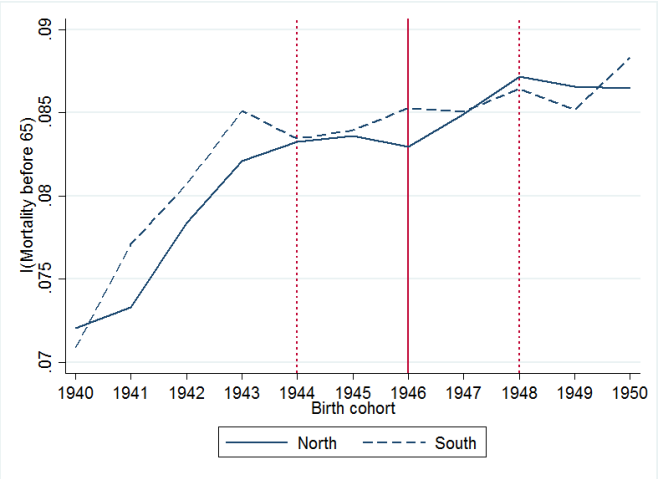


(b) Earnings in 1999 (ln)

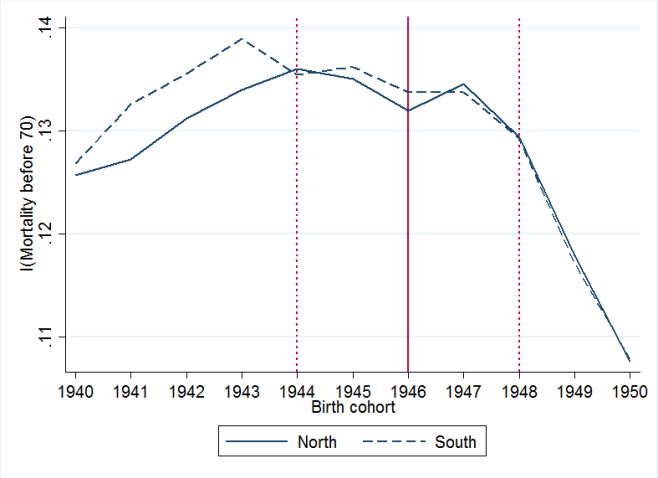


(c) Social security benefits in 1999 (0/1)

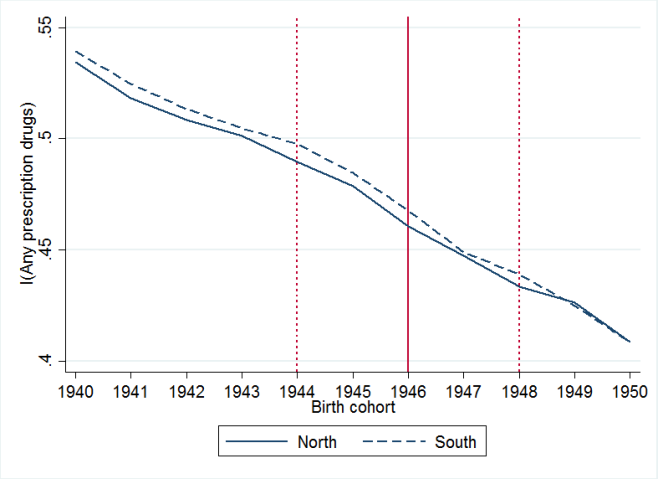
Figure 2.A9: Common trend in labor market outcomes



(a) Mortality before age 65 (0/1)



(b) Mortality before age 70 (0/1)



(c) Any prescription drugs, four groups related to lifestyle (0/1)

Figure 2.A10: Common trend in health outcomes

Appendix B: Additional Tables

Table 2.A1: Calorie rations by regions by three months. Source: (Stein et al., 1975)

Area	June-Aug 1944	Sept-Nov 1944	Dec-Feb 1944-45	Mar-May 1945	June-Aug 1945	Sept-Nov 1945
West	1512	1414	740	670	1757	2083
North	1512	1450	1345	1392	1755	2083
South	1512	1403	1375	1692	1864	2083

Table 2.A2: Maternal age at birth in 1946 and age in 1995.

Maternal age in 1946	Maternal birth year	Age in 1995
20	1926	69
21	1925	70
22	1924	71
23	1923	72
24	1922	73
25	1921	74
26	1920	75
27	1919	76
28	1918	77
29	1917	78
30	1916	79
31	1915	80
32	1914	81
33	1913	82
34	1912	83
35	1911	84

Table 2.A3: Common trend assumption tests - parental level

	Maternal age at birth (1)	Paternal age at birth (2)	Length of marriage (3)
D41	-0.116	-0.139	27.357
	(0.096)	(0.103)	(92.210)
D42	0.087	0.067	28.236
	(0.088)	(0.084)	(72.995)
D43	0.029	0.058	77.690
	(0.076)	(0.085)	(67.875)
D44	0.096	0.131	46.760
	(0.089)	(0.090)	(67.563)
D45	0.096	0.172*	109.318
	(0.084)	(0.088)	(69.607)
D46	0.080	0.072	136.075*
	(0.068)	(0.075)	(69.978)
D47	0.091	0.106	65.996
	(0.077)	(0.082)	(66.947)
D48	0.067	0.125*	99.078
	(0.068)	(0.074)	(65.142)
D49	0.055	0.166**	78.647
	(0.069)	(0.077)	(63.945)
D50	0.041	0.159	84.526
	(0.064)	(0.070)	(64.725)
N	1,035,910	1,031,995	873,489

Notes : Estimated by OLS. The outcome of interest is regressed on a series of region and year fixed effects (not reported). The reported coefficients represent interactions between birth year and being born in the north. Standard errors are clustered by birthmonth/year and region.

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

Table 2.A4: Common trend assumption tests - child level

	Employed	Earnings (ln)	Any Social Security	Mort. ≤ 65	Mort. ≤ 70	Any drugs 4 groups
	(1)	(2)	(3)	(4)	(5)	(6)
D41	-0.008	-0.096	-0.001	-0.005	-0.004	-0.001
	(0.012)	(0.199)	(0.010)	(0.004)	(0.004)	(0.007)
D42	-0.011	-0.134	0.005	-0.003	-0.003	-0.000
	(0.011)	(0.174)	(0.009)	(0.004)	(0.004)	(0.007)
D43	-0.018*	-0.244	0.007	-0.004	-0.004	0.001
	(0.011)	(0.166)	(0.010)	(0.003)	(0.005)	(0.007)
D44	-0.017*	-0.261*	0.004	-0.001	0.002	-0.003
	(0.010)	(0.151)	(0.008)	(0.004)	(0.004)	(0.007)
D45	-0.028***	-0.413***	0.019**	-0.001	-0.000	-0.001
	(0.010)	(0.153)	(0.009)	(0.004)	(0.004)	(0.008)
D46	-0.024**	-0.333**	0.024***	-0.003	-0.001	-0.002
	(0.009)	(0.146)	(0.008)	(0.003)	(0.004)	(0.007)
D47	-0.027***	-0.359**	0.020**	-0.001	0.002	0.003
	(0.009)	(0.146)	(0.008)	(0.003)	(0.004)	(0.006)
D48	-0.031***	-0.444***	0.023***	-0.000	0.001	-0.001
	(0.010)	(0.150)	(0.008)	(0.003)	(0.004)	(0.007)
D49	-0.035***	-0.486***	0.031***	0.000	0.002	0.006
	(0.009)	(0.145)	(0.007)	(0.003)	(0.004)	(0.006)
D50	-0.031***	-0.471***	0.028***	-0.003	0.001	0.005
	(0.009)	(0.145)	(0.007)	(0.003)	(0.004)	(0.006)
N	530,117	530,117	530,117	1,035,910	1,035,910	1,035,910

Notes : Estimated by OLS. The outcome of interest is regressed on a series of region and year fixed effects (not reported). The reported coefficients represent interactions between birth year and being born in the north. Standard errors are clustered by birthmonth/year and region.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.A5: Log earnings conditional on employment

	Labor earnings (ln)			
	Pre-sample		Post-sample	
	(1)	(2)	(3)	(4)
Panel A				
BP	0.020**	-0.070***	-0.009	0.012
	(0.008)	(0.026)	(0.006)	(0.017)
North	0.018**	0.004	0.010**	-0.036
	(0.007)	(0.021)	(0.005)	(0.028)
CoBP	0.003	0.026	0.011	0.021
	(0.010)	(0.024)	(0.008)	(0.017)
Controls	No	Yes	No	Yes
Mean dep. var.	10.281	10.281	10.298	10.298
N	112,484	112,484	149,743	149,743
Panel B: By marital status at conception				
BP	-0.045	-0.160	0.021	0.102
	(0.037)	(0.104)	(0.027)	(0.073)
North	-0.003	0.151*	-0.103*	0.116
	(0.046)	(0.077)	(0.052)	(0.141)
CoBP	-0.003	0.030	0.066**	-0.064
	(0.033)	(0.089)	(0.032)	(0.076)
Controls	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG
Mean dep. var.	10.362	10.249	10.381	10.245
N	27,623	6,098	37,646	8,539

Notes: Estimated by OLS. Standard errors are clustered by birth month/year and birth region (north versus south). The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and a control for maternal age at birth. The pre-sample specification also contains an indicator for being conceived during the liberation of the south.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.A6: Results - Other health outcomes

	Any drugs	Mental	Cardiovas.	Respir.	Diabetes
Panel A: Pre-sample					
BP	0.030***	0.024***	0.019**	0.009	-0.008
	(0.008)	(0.007)	(0.008)	(0.007)	(0.005)
North	-0.007	0.015***	-0.013**	-0.003	-0.004
	(0.006)	(0.005)	(0.006)	(0.004)	(0.003)
CoBP	-0.020***	-0.014**	-0.014*	-0.004	0.008*
	(0.007)	(0.006)	(0.008)	(0.006)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes
Mean dep. var.	0.744	0.191	0.330	0.086	0.067
N	267,380	267,380	267,380	267,380	267,380
Panel B: Pre-sample for first births conceived in-wedlock					
BP	0.049***	0.031**	0.018	0.001	0.009
	(0.013)	(0.013)	(0.019)	(0.011)	(0.012)
North	0.003	0.018**	-0.024	-0.006	0.000
	(0.011)	(0.008)	(0.016)	(0.008)	(0.007)
CoBP	-0.028*	-0.018	-0.017	-0.018	-0.005
	(0.015)	(0.013)	(0.019)	(0.011)	(0.011)
Controls	Yes	Yes	Yes	Yes	Yes
Mean dep. var.	0.742	0.189	0.321	0.083	0.062
N	65,085	65,085	65,085	65,085	65,085
Panel A: Pre-sample for first births conceived out of wedlock					
BP	0.025	-0.000	-0.022	0.008	-0.041
	(0.037)	(0.024)	(0.037)	(0.034)	(0.026)
North	-0.037*	-0.001	-0.060**	-0.033**	0.003
	(0.021)	(0.017)	(0.025)	(0.014)	(0.016)
CoBP	-0.052	0.003	0.003	0.023	0.021
	(0.032)	(0.022)	(0.035)	(0.031)	(0.024)
Controls	Yes	Yes	Yes	Yes	Yes
Mean dep. var.	0.746	0.188	0.358	0.094	0.084
N	14,390	14,390	14,390	14,390	14,390

Notes : Any drugs corresponds to any prescription for any drugs and not, as earlier only for the relevant four groups. Double-difference estimate Birth'46*North is reported, full control specification.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.A7: Family characteristics - Post-sample

	Family size		Marriage stability	
	(1)	(2)	(3)	(4)
Panel A: Post-sample				
BP	-0.085***	0.078	65.56*	12.10
	(0.026)	(0.053)	(36.04)	(88.29)
North	-0.382***	-0.480***	259.53***	388.72**
	(0.023)	(0.093)	(22.77)	(173.44)
CoBP	-0.131***	-0.036	39.88	-24.78
	(0.043)	(0.055)	(47.20)	(112.83)
Controls	No	Yes	No	Yes
Mean dep. var.	4.412	4.412	16,705.47	16,705.47
N	342,551	342,551	294,189	294,189
Panel B: By marital status at conception				
BP	0.019	-0.279	8.83	660.10
	(0.083)	(0.188)	(183.37)	(427.56)
North	-0.583***	-0.506*	314.06	1221.63*
	(0.183)	(0.269)	(319.36)	(706.68)
CoBP	-0.101	0.274	8.09	-813.22*
	(0.102)	(0.189)	(199.06)	(424.49)
Controls	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG
Mean dep. var.	3.366	3.581	16,441.27	16,505.49
N	85,087	19,474	85,087	19,474

Notes: Estimated by OLS. Standard errors are clustered by birth month/year and birth region (north versus south). The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and a control for maternal age at birth.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.A8: Labor market outcomes - Males - Post-sample

	Employment		Labor earnings (ln)		Social security	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Post-sample						
BP	-0.017***	0.003	-0.267***	0.071	0.012***	-0.013*
	(0.003)	(0.007)	(0.053)	(0.121)	(0.004)	(0.007)
North	0.018***	0.034**	0.274***	0.491**	-0.018***	-0.027*
	(0.003)	(0.013)	(0.050)	(0.213)	(0.004)	(0.014)
CoBP	-0.000	-0.012	-0.004	-0.164	0.007	0.014*
	(0.004)	(0.008)	(0.067)	(0.130)	(0.006)	(0.008)
Controls	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.865	0.865	8.122	8.122	0.204	0.204
N	175,354	175,354	175,354	175,354	175,354	175,354
Panel B: Post-sample by marital status at conception						
BP	-0.004	0.023	-0.070	0.524	-0.018	-0.082*
	(0.015)	(0.041)	(0.238)	(0.642)	(0.014)	(0.043)
North	0.035	0.127**	0.459	2.222**	-0.021	-0.136*
	(0.027)	(0.061)	(0.409)	(1.002)	(0.024)	(0.079)
CoBP	-0.020	-0.035	-0.291	-0.693	0.015	0.085*
	(0.018)	(0.041)	(0.274)	(0.656)	(0.015)	(0.045)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.878	0.851	8.381	7.927	0.166	0.221
N	43,444	10,119	43,444	10,119	43,444	10,119

Notes: Estimated by OLS, males only, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and controls for whether the individuals is born in a city, and maternal age at birth. IW indicates that the individual is conceived in-wedlock, whereas SG denotes a shotgun conception.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.A9: Health outcomes - Post-sample

	Mortality ≤ 65		Mortality ≤ 70		Any drugs (4)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Post-sample						
BP	0.000	-0.003	0.004	0.004	0.030***	0.007
	(0.002)	(0.003)	(0.003)	(0.006)	(0.005)	(0.007)
North	-0.000	-0.008	0.001	-0.002	-0.003	-0.010
	(0.001)	(0.006)	(0.001)	(0.008)	(0.003)	(0.012)
CoBP	-0.003	0.002	-0.005	-0.003	-0.008	-0.005
	(0.002)	(0.003)	(0.003)	(0.005)	(0.006)	(0.007)
Controls	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.085	0.085	0.132	0.132	0.448	0.448
N	342,551	342,551	342,551	342,551	342,551	342,551
Panel B: Post-sample by marital status at conception						
BP	0.003	-0.022	0.006	-0.025	-0.006	-0.040
	(0.007)	(0.014)	(0.011)	(0.024)	(0.017)	(0.034)
North	0.006	0.003	0.032	-0.033	-0.026	-0.123**
	(0.013)	(0.024)	(0.021)	(0.041)	(0.031)	(0.052)
CoBP	-0.009	0.008	-0.023*	0.020	-0.005	0.047
	(0.008)	(0.015)	(0.012)	(0.024)	(0.018)	(0.033)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.079	0.095	0.124	0.145	0.442	0.482
N	85,087	19,474	85,087	19,474	85,087	19,474

Notes: Estimated by OLS, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and controls for whether the individuals is born in a city, and maternal age at birth. IW indicates that the individual is conceived in-wedlock, whereas SG denotes a shotgun conception.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.A10: Labor market outcomes - Cohort size controls

	Employment		Labor earnings (ln)		Social security	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pre-sample						
BP	0.005	0.004	0.029	0.018	0.002	0.004
	(0.010)	(0.011)	(0.165)	(0.169)	(0.013)	(0.013)
North	0.032***	0.031***	0.459***	0.431***	-0.039***	-0.035***
	(0.007)	(0.007)	(0.115)	(0.116)	(0.006)	(0.006)
CoBP	-0.007	-0.010	-0.117	-0.177	0.003	0.010
	(0.009)	(0.009)	(0.129)	(0.135)	(0.012)	(0.012)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Coh. size controls	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.834	0.834	7.644	7.644	0.237	0.237
N	136,698	136,698	136,698	136,698	136,698	136,698
Panel B: Pre-sample by marital status at conception						
BP	-0.031	0.035	-0.473	0.185	0.016	0.047
	(0.023)	(0.038)	(0.359)	(0.583)	(0.019)	(0.064)
North	0.027***	0.065*	0.411**	0.805	-0.025**	-0.083*
	(0.010)	(0.033)	(0.163)	(0.531)	(0.011)	(0.045)
CoBP	-0.007	-0.013	-0.242	0.092	-0.002	-0.050
	(0.021)	(0.032)	(0.332)	(0.493)	(0.018)	(0.051)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Coh. size controls	Yes	Yes	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.844	0.829	7.853	7.567	0.201	0.238
N	33,186	7,442	33,186	7,442	33,186	7,442

Notes: Estimated by OLS, males only, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and controls for whether the individuals is born in a city, and maternal age at birth. The pre-sample specification also contains an indicator for being conceived during the liberation of the south. IW indicates that the individual is conceived in-wedlock, whereas SG denotes a shotgun conception.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.A11: Health outcomes - Cohort size controls

	Mortality ≤ 65		Mortality ≤ 70		I(Lifestyle drugs)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pre-sample						
BP	0.001	0.001	0.003	0.003	0.035***	0.035***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
North	0.000	0.000	0.000	0.001	-0.003	-0.002
	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)
CoBP	-0.002	-0.002	-0.005	-0.004	-0.024***	-0.022***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Coh. size controls	No	Yes	No	Yes	No	Yes
Mean dep. var.	0.084	0.084	0.135	0.135	0.479	0.479
N	267,380	267,380	267,380	267,380	267,380	267,380
Panel B: Pre-sample by marital status at conception						
BP	-0.011	-0.019	-0.013	0.012	0.044***	0.059
	(0.011)	(0.032)	(0.011)	(0.043)	(0.016)	(0.045)
North	-0.005	0.008	-0.007	-0.016	-0.005	-0.037
	(0.008)	(0.017)	(0.008)	(0.029)	(0.014)	(0.023)
CoBP	0.011	0.032	0.009	0.008	-0.048***	-0.071*
	(0.010)	(0.027)	(0.011)	(0.036)	(0.016)	(0.041)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Coh. size controls	Yes	Yes	Yes	Yes	Yes	Yes
Group	IW	SG	IW	SG	IW	SG
Mean dep. var.	0.081	0.085	0.129	0.139	0.475	0.502
N	65,085	14,390	65,085	14,390	65,085	14,390

Notes: Estimated by OLS, restricted sample. The specification contains region-specific birth-month/year trends (linear and quadratic), birth month dummies, and controls for whether the individuals is born in a city, and maternal age at birth. The pre-sample specification also contains an indicator for being conceived during the liberation of the south. IW indicates that the individual is conceived in-wedlock, whereas SG denotes a shotgun conception.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Chapter 3

The role of prenatal testosterone in test scores[1]

Joint work with Anne C. Gielen

3.1 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

[1]This chapter is based on Gielen and Zwiers (2018). The authors wish to thank the Centraal Instituut voor Toestsontwikkeling (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.

(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.[2] 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).[3] 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 characteris-

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

[3]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).

tics, exposure to prenatal testosterone is known to wire the brain with masculine behavioral patterns (i.e. in preferences, personality, and temperament) (Jordan-Young, 2010).[4] 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 exploiting 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).[5] Since these male-typical cognitive capabilities, e.g. spatial skills,

[4]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.

[5]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).

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.[6] 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 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.

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

3.2 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. Knudsen et al., 2006; Heckman, 2008; Almond and Currie, 2011a). 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, 2011b; Scholte et al., 2015). This paper considers the impact of prenatal exposure to testosterone on educational performance in childhood.

3.2.1 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 lifecycle. 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).[7] 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).[8] 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.

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 diag-

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

[8] 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.

nosed with CAH are found to perform better on spatial tasks than control women (Puts et al., 2008). Disadvantages of using clinical samples are the usually small sample 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).

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.

3.3 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.[9]

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 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

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

twins[10], 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).[11] To control for this, we define a control group of closely spaced singletons (CSS), consisting of singletons who have a sibling

[10]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).

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

born within 12 months of their own birth date.[12] Provided that any sibling socialization effects are similar for twins and for singletons in the CSS sample[13], 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). Although 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

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

[13]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.

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 3.1, 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.

$$\begin{aligned} y_{it} = \beta_0 + \beta_1 female_i + \beta_2 OS_i + \\ \beta_3 twin_i + \beta_4(female_i * OS_i) + \beta_5(twin_i * female_i) + \\ \beta_6(twin_i * OS_i) + \beta_7(female_i * OS_i * Twin_i) + \mathbf{X}_{it}\delta + u_{it} \end{aligned} \tag{3.1}$$

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.

3.4 Data

3.4.1 Dutch twins

This paper uses administrative data from Statistics Netherlands covering all registered inhabitants of the Netherlands.[14] 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 struc-

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

tures is shown in Table 3.1. The twinning probability (3.26%) is consistent with the incidence of twinning in the Netherlands between 1993 and 2004 (3.39%).[15] We proceed with a sample of twins and CSS, dropping singletons without siblings, singletons with siblings born outside the 12 month range, and higher order multiples.

Table 3.1: 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 3.2.[16]

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

[16]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.

Table 3.2: 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.

3.4.2 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.[17] The data cover the years 2006 to 2014.[18] For children having multiple test-score records in this period (e.g. due to class retention) the most recent score is preserved.

[17]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.

[18]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.

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 3.2.

The standardized test incorporates performance measures for language, math, information processing, and world orientation.[19] The scores on the various (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.[20]

3.4.3 Descriptive statistics

Average standardized test scores differ between boys and girls, and between twins and CSS (Table 3.3).[21] 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 3.4 shows that gender gaps in test performance also vary with the gender of one's sibling.[22] Although we observe no significant differences in test scores between opposite-sex and same-sex closely spaced sin-

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

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

[21]Exact variable definitions are provided in Appendix Table 3.A1.

[22]Gender differences in the distribution of test scores are presented in Figure 3.A1 and Figure 3.A2.

Table 3.3: 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

gletons, 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 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[23], 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 technolo-

[23]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.

Table 3.4: 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 3.

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

gies (ART) (Bhalotra et al., 2016), we observe that twins are born to older mothers (and fathers) in high income families.[24] 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.

3.5 Results

The results from the baseline specification for the aggregate test score are presented in Table 3.5. The twin coefficient is positive and significant in the specification without controls (column 1), and becomes smaller and insignificant once 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,

[24]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 3.5: 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

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

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 3.6, respectively. Twins appears to have higher math and reading scores than CSS, but these differences disappear once we

Table 3.6: 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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.[25]

The results in Tables 3.5 and 3.6 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 3.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 3.6 we discuss various explanations for our findings.

[25]Table 3.A2 and Table 3.A3 show that the results are robust to estimating the models separately for boys and girls.

3.5.1 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 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 4.9 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.[26,27]

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.[28]

Table 3.4 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.[29] 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 4.9 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.[30] Hence, these estimates confirm our main results and, if anything, may suggest that our baseline estimate is somewhat

[26]The full estimation results are available in Appendix Tables 3.A4 and 3.A5.

[27]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).

[28]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.).

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

[30]The full estimation results are available in Appendix Tables 3.A4 and 3.A5.

Table 3.7: 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 3.A4 and 3.A5.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

conservative.

Another difference between families with CSS and families with twins is the number of children in a household (see Table 3.4). 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 fam-

ilies only such that twins and CSS in the sample grow up in families of equal size (Panel D of Table 4.9). The double difference estimates 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 4.9 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 3.A8.[31] 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[32], 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.

3.6 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

[31]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).

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

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 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 3.8 shows the results from a model in which we look at birth weight as the relevant outcome.[33] 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,

[33]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.

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.

Table 3.8: 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 3.A9
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 3.9 shows, however, that opposite-sex twinning does not seem to be related with any measure of grade retention.[34] Furthermore, a poten-

[34]We use two proxies for grade retention as a direct measure is unavailable. We use

Table 3.9: 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-	0.011	-0.012
	(0.003)	(0.002)	vocational track	(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-	0.007	-0.004
	(0.005)	(0.005)	vocational track	(0.013)	(0.013)
N	43,069	43,069	N	30,944	30,944
			- At least general/higher	0.009	0.002
			pre-vocational track	(0.016)	(0.017)
			N	30,944	30,944
			- At least higher pre-	0.013	-0.003
			vocational track	(0.017)	(0.017)
			N	30,944	30,944
			- At least higher pre-	-0.018	-0.006
			vocational/general track	(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).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

tial differential impact of non-cognitive skills on overall school performance is also not reflected in the teachers' assessment of the student's overal-

an indicator for having multiple Cito records in the data for retention in the final grade of elementary school. We proxy any grade retention with an indicator variable for being 13 years or older at the time of taking the test.

l ability.[35] 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.

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 (2018) find that family disadvantage disproportionally negatively affects the behavior and school outcomes of boys relative to girls. In Table 3.10 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

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

Table 3.10: 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 3.A10, 3.A11, and 3.A12.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

depend on the environment in which a child is being raised.[36] Table 3.10 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 breadwinner 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 3.10 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

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

(Reardon et al., 2018). However, to get a better understanding of this mechanism we exploit regional differences in traditional gender norms across 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.[37] 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).[38] 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 3.11 shows that the double-difference estimators are more negative for girls living in more religious areas.[39] 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

[37]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.

[38]See column 5 and 6 of Table 3.A13.

[39]The full estimation results are available in Table 3.A13.

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.

Table 3.11: 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.7 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 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.

Appendix

Table 3.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 3.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

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 3.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

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

	Aggregate score			
	(1) Baseline	(2) First born	(3) Native kids	(4) 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 3.A5: Robustness results for standardized reading and math score - sub-sample estimates

	Reading score				Math score			
	(1) Baseline	(2) First born	(3) Native kids	(4) Two-child families	(5) Baseline	(6) First born	(7) Native kids	(8) 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 3.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 3.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 3.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 3.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

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 3.A10: Mechanisms at work - standardized aggregate score

	Aggregate score									
	(1) Baseline	(2) Two parent HH	(3) One parent HH	(4) Non-divorced	(5) Divorced	(6) Not married	(7) Low-income	(8) High-income	(9) Dis-advantaged	(10) 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 3.A11: Mechanisms a work - standardized reading score

	Reading score									
	(1) Baseline	(2) Two parent HH	(3) One parent HH	(4) Non-divorced	(5) Divorced	(6) Not married	(7) Low-income	(8) High-income	(9) Dis-advantaged	(10) 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 3.A12: Mechanisms at work - standardized math score

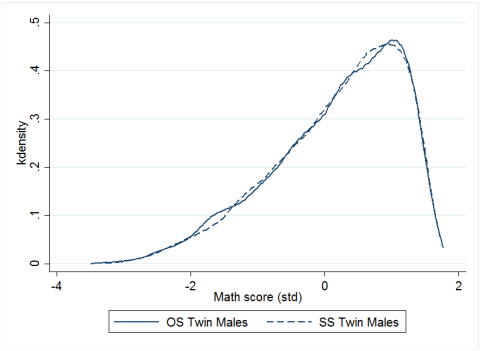
	Aggregate score									
	(1) Baseline	(2) Two parent HH	(3) One parent HH	(4) Non-divorced	(5) Divorced	(6) Not married	(7) Low-income	(8) High-income	(9) Dis-advantaged	(10) Advantaged
Twin	-0.036	-0.022	-0.097*	0.001	-0.225***	-0.046	-0.065**	0.040	-0.152***	-0.025
	(0.023)	(0.025)	(0.057)	(0.026)	(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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

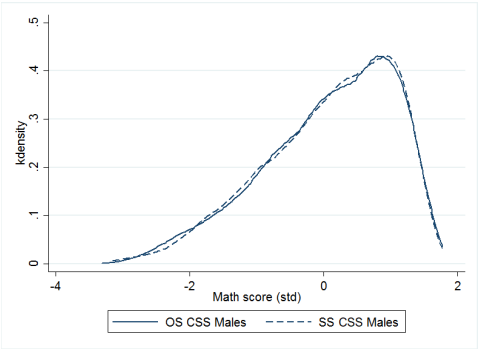
Table 3.A13: By traditional family norms

	Aggregate score		Reading score		Math score	
	(1)	(2)	(3)	(4)	(5)	(6)
	Less Religious	More Religious	Less Religious	More Religious	Less Religious	More 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

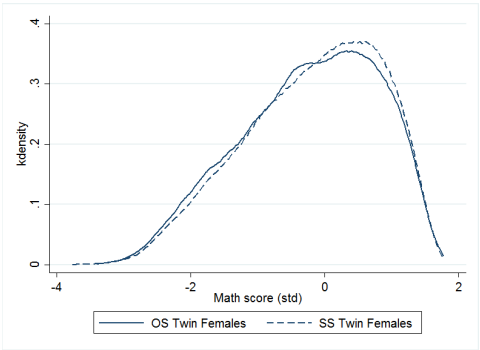
Notes: Results are based on OLS model. The set of controls is similar to that in Table 5. Standard errors are clustered on maternal ID and are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.



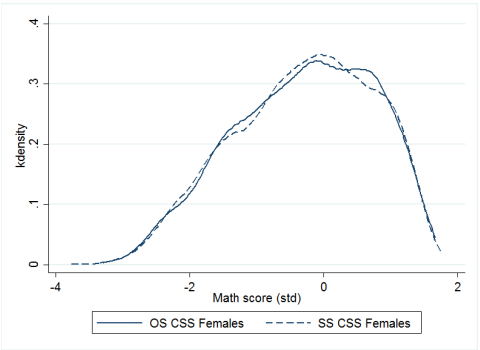
(a) Male Twins



(b) Male CSS

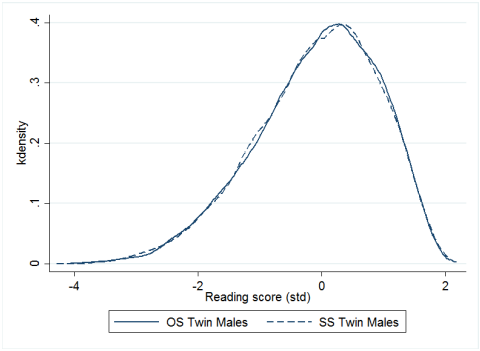


(c) Female Twins

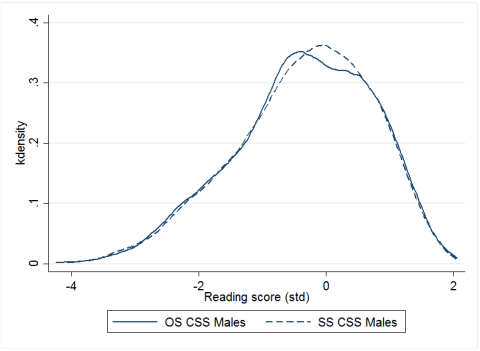


(d) Female CSS

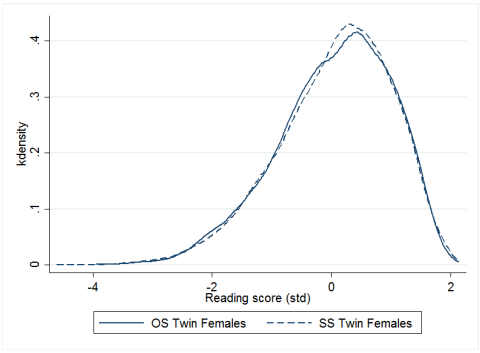
Figure 3.A1: Test-score distributions for math, by gender, sibling-type and sibling gender.



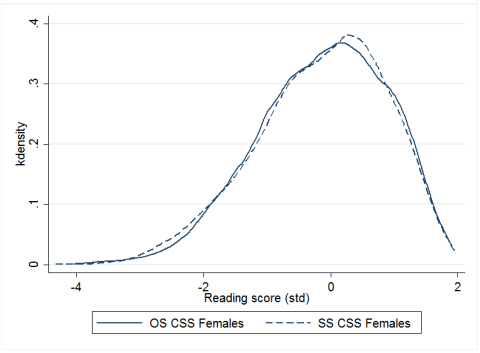
(a) Male Twins



(b) Male CSS

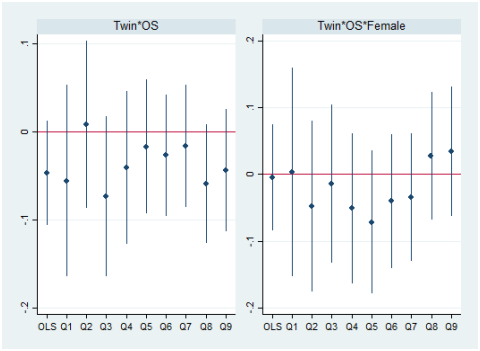


(c) Female Twins

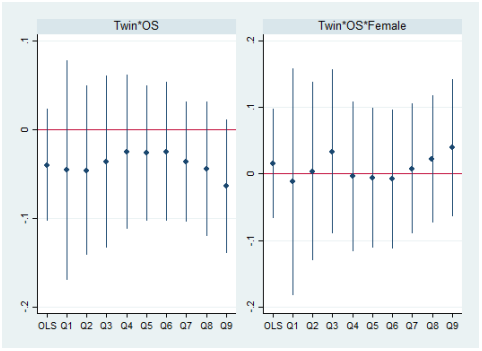


(d) Female CSS

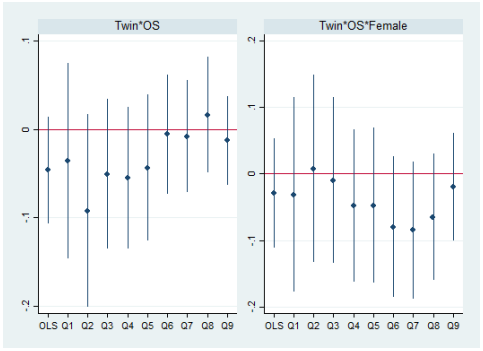
Figure 3.A2: Test-score distributions for reading, by gender, sibling-type and sibling gender.



(a) Total score



(b) Reading score



(c) Math score

Figure 3.A3: Quantile regression, and 95% confidence interval

Chapter 4

Last but (not) least? Aversion to the lowest educational track[1]

4.1 Introduction

When it comes to perceptions about several socioeconomic outcomes, individuals tend to not only care about their position in absolute terms, but also about their position relative to others. To illustrate, conditional on the individual's own earnings, the income of others in the neighborhood affects wellbeing negatively (Blanchflower and Oswald, 2004; Luttmer, 2005). When employees learn that they earn less than comparable colleagues, job satisfaction is lower (Card et al., 2012). The literature tends to focus on the effects of relative income on utility, but these relative effects may also play a role in other life outcomes. Very little is known about potential non-linearities across the distribution, and the findings in the literature diverge. Blanchflower and Oswald find that effects are strongest for the top of the

[1]I wish to thank Anne Gielen and Dinand Webbink for useful discussions and support. This paper benefited from comments made by Lammertjan Dam, David Figlio, Aenneli Houkes-Hommes, Sacha Kapoor, Kevin Lang, Olivier Marie, Jesse Rothstein, Ulf Zölitz, and seminar participants at Erasmus School of Economics, CPB Netherlands Bureau for Economic Policy Analysis, and the 28th EALE Conference (Ghent). All errors and emissions are my own.

distribution, Card et al. find stronger effects for those in the bottom, and Luttmer finds no differences across the distribution. Kuziemko et al. (2014) focus on one type of non-linearity, namely: last place aversion. Last place aversion refers to a setting in which individuals have a particular dislike for being in last place. Whereas Kuziemko et al. examine distributional preferences in income, I study last place aversion in an educational context by exploiting a reform that exogenously changed the lowest track in Dutch secondary education. Last place aversion may have persistent consequences in this setting as of the importance of ability tracking for later life outcomes (e.g. Betts, 2011).

This paper studies last place aversion in an educational context, and specifically focuses on a reform that changed the lowest tracks in Dutch secondary education. The reform was introduced in 1999 and merged the then two lowest secondary school tracks. The second-to-lowest track suddenly became the lowest track, and hence this represents a unique setting to study last place aversion. In the Netherlands, children are tracked after 6^{th} grade when they are about age twelve. The outcomes studied in this paper are the variables that determine track allocation in 6^{th} grade: (1) the child's performance on a high-stakes standardized cognitive test, and (2) a track recommendation made by the primary school teacher.

These outcomes are expected to change after the reform if last place aversion plays a role. Students and parents can influence track allocation in secondary school by affecting the two variables that determine track allocation. To illustrate, parents can invest in homework assistance to improve the child's test performance, children may experience increased test motivation, and parents may put pressure on primary school teachers in

order for them to give a higher track recommendation. Anecdotal evidence in newspaper articles suggest that these types of behaviors take place after the introduction of the new lowest track. The Dutch Inspectorate of Education reports that about 52% of schools experience parental pressure when determining the track recommendation, of which 44% is aimed at receiving higher track recommendations. Last place aversion to the new lowest track is mentioned as one of the explanations for this finding (Inspectie van het Onderwijs, 2014).

To study if the reform led to last place aversion, and consequently to differences in the variables that determine track allocation, I employ a difference-in-difference strategy. I combine the timing of the reform with the fact that only children who are expected to attend the lowest track are affected by the reform. The data comes from the longitudinal PRIMA survey, which was administered biannually from 1994 to 2005. It is repeated cross-sectional data with three waves of data before the reform and three years of data after the reform. The dataset contains information on a low-stakes standardized cognitive test that is not important for track allocation in secondary school. This test-score is used to determine which children are likely to attend the lowest track, and hence are in the treatment group. A treatment group with children who are likely to attend the lowest track based on their reading and math score is established. The control group consists of children who are just above the bar.

The results of the difference-in-difference approach show that children who are likely to attend the lowest track have a significantly lower chance of receiving a track recommendation higher than the lowest track. The effect is particularly large for those in the treatment group based on their math

score: a 6.5 percentage points lower probability compared to a mean enrollment of 44.5%. Those who are expected to attend the lowest track based on their math score also have a significantly lower score on the high-stakes standardized cognitive test after the reform (lower by 5% of a standard deviation). No significant effects are found for children who are expected to attend the lowest track based on their reading score. If anything, they have higher test-scores after the reform.

I examine what mechanisms can explain the results found. First, several types of behavioral responses could lead to differences in the variables that determine track allocation, ranging from indirectly trying to influence the high-stakes standardized test-scores to directly influencing the teacher for a higher track recommendation. To examine the importance of these channels I add a control for the high-stakes test score to the specification with the track recommendation as the relevant outcome. The results suggest that lower track recommendations cannot be explained by lower test-scores.

Second, I examine what is driving the lower test-scores after the reform for children who are expected to attend the lowest track. The main analysis is performed on the aggregate score, but separate scores for math and reading are available for a sub-sample. What stands out is that the effects are stronger for the child's weakest subject. To illustrate: if a child is bad at math, he or she scores lower at math after the reform. The pressure to perform well might be higher for the child's weakest subject, which may explain why scores are lower.

Finally, I find that the effects are larger, i.e. more negative, among children in families with higher-educated parents. Literature shows that higher-educated parents have a higher valuation of education (Hastings

et al., 2005; Burgess et al., 2015), are more aware of differences in the quality of education (Sacerdote et al., 2011; Borghans et al., 2015), prefer grade retention over transition to a lower track (Kloosterman and de Graaf, 2010), and invest more in homework assistance (Rønning, 2011). I hypothesize that a feeling of last place aversion is stronger in these families, and hence a higher pressure to perform well may lead to lower test-scores.

This paper adds to the literature in several ways. First, I add to the literature on how individual wellbeing is affected by their position relative to others (Blanchflower and Oswald, 2004; Luttmer, 2005; Card et al., 2012), but in particular to the literature on last place aversion. Kuziemko et al. (2014) provide evidence for last place aversion in a laboratory setting. In a first experiment they show that participants are more likely to choose gambles that let them move up in the income distribution if they are in last place as opposed to other places in the distribution. In a second experiment they show that in a modified dictator game participants in second-to-last place are more likely to give money to the person ranked above them than to the person ranked below to preserve their rank. Finally, they show that individuals earning slightly above minimum wage are more likely to oppose minimum wage increases, and individuals earning below the median but above poverty are more likely to oppose redistribution compared to what their background characteristics would suggest. Both distributional preferences are consistent with last place aversion. There is little evidence on last place aversion, and this is to the best of my knowledge the first application to an educational context. My findings suggest that last place aversion does not yield higher test-scores or higher track recommendations. If anything, children do worse, and I find suggestive evidence that children

perform especially worse if the pressure is higher.

Second, I add to the literature on the influence of parental inputs on child performance (Figlio and Lucas, 2004; Houtenville and Conway, 2008; Rønning, 2011). This paper adds especially to the literature that examines interactions between parental inputs and external factors. To illustrate, parental involvement in the child's education increases with the introduction of Head Start (Gelber and Isen, 2013), decreases with increases in school resources (Houtenville and Conway, 2008), and increases if teachers employ tougher grading (Figlio and Lucas, 2004). This paper adds to this literature by studying how changes in the institutional setting of an educational system affect parental behaviors, which may affect the child's educational outcomes.

Third, I add to the literature on the potential effects of tracking. The literature shows that tracking exacerbates existing differences between children (see for an overview Betts, 2011). However, very little is known about potential unintended effects of tracked educational systems. These unintended effects can arise if parents and/or children want to avoid the lowest educational track. This may lead to a potential mismatch between the ability of the child and the educational track he is enrolled. This paper shows that last place aversion leads to lower track recommendations and lower test scores for those expected to attend the lowest track. Given the persistent effects of ability tracking, this may have long-run consequences.

The structure of the paper is as follows. The next Section outlines the institutional setting of the reform. The empirical strategy and the data are discussed in Section 3 and 4 respectively. The results are provided in Section 5, and Section 6 explores several mechanisms that could explain

the results found. This paper ends with a conclusion in Section 7.

4.2 Background and the reform

4.2.1 Tracking in the Netherlands

Within the Dutch educational system children attend eight years of primary education starting at age four. The first two years of primary school are kindergarten, and children generally start learning to read and write in the third year of primary school, which is equivalent to first grade (Feron et al., 2016). After finishing primary education, children are tracked into different educational levels in secondary school at approximately age twelve (transition from 6^{th} grade to 7^{th} grade).[2] Secondary education takes four to six years depending on the level of the track. In the studied time period, track allocation in secondary school is based on two variables that are determined by the end of primary school: (1) the child's score on a high-stakes standardized cognitive test taken in 6^{th} grade, and (2) a track recommendation by the primary school teacher.

The standardized cognitive test that is important for track allocation is better known as the Cito-test. Administering the Cito-test is not mandatory for schools, but more than 80% of Dutch schools administer the test (Chorny et al., 2010). The test is administered in 6^{th} grade in February, and the test procedure is similar for all children.[3] The test covers four main areas: reading, math, information processing and world orientation (the latter is optional), and children can obtain scores ranging from 501

[2]This is at a younger age than in other OECD countries, which usually track children starting from age 14-16 (Feron et al., 2016).

[3]The test is developed by the Centraal Instituut for Toetsontwikkeling (CITO). The test is no longer administered in February from 2015.

to 550. The test is very important as the test-scores are used as input by the primary school teacher when determining their track recommendation, and secondary schools can set score-thresholds as to determine which child is allowed in which track.[4,5]

The track recommendation made by the teacher is based on their experience with the student and their evaluation of the child's ability. It is not only based on the child's performance in the 6^{th} grade, but also in earlier grades. Observable family and socioeconomic characteristics may also influence the teacher recommendation (Feron et al., 2016). Teachers know the standardized cognitive test score at the time of determining the appropriate track recommendation for each child. The final track allocation decision is made by the secondary school the child will attend starting from 7^{th} grade. This decision is based on both the cognitive test score as well as the teacher's track recommendation.

4.2.2 The reform

Before the reform in 1999, Dutch secondary education consisted out of four tracks (Table 4.1). An academic track (abbreviated *in Dutch*: vwo) that prepared students for university education in six years. A general track (havo) that prepared students for higher vocational education in five years. General secondary education (mavo) that prepared students for vocational education in four years, and lower vocational education (lbo) that prepared students for vocational education in four years. The two lowest tracks both prepared students for vocational education, although lower vocational ed-

[4]The scores are also important for schools as they are used to measure school quality (Chorny et al., 2010; Feron et al., 2016).

[5]A guideline by Van Boxtel et al. (2011) for these track allocation based on the Cito cognitive test score is provided in Table 4.A1.

ucation had a more practical focus as compared to general vocational education. There are two tracks for special needs students, which are separate from the main system. I list these tracks as special (ivbo) and special+ (vso), where children receive more support when they are enrolled in the latter track.

This system changed with the implementation of a reform in 1999 as is shown in Table 4.1. The reform combined the previous two lowest tracks in secondary education, i.e. general vocational education (mavo) and lower vocational education (lbo) into the pre-vocational track (vmbo). The reform left the academic track and the general track unaffected. The newly established pre-vocational track prepares students for vocational education in four years. The main goal of the reform was to improve the transition from pre-vocational education to vocational education and the labor market (Ministry of Education, 2005).

The newly introduced pre-vocational track again consists out of four sub-tracks that are different in level and practical orientation. The theoretical (vmbo-tl) and mixed track (vmbo-gl) are comparable to what was previously known as the general vocational track (mavo). The middle management (vmbo-kbl) and basic track (vmbo-bbl) are comparable to what was previously known as the lower vocational track (lbo). The ordering of the sub-tracks in Table 4.1 corresponds to tracks with a more theoretical focus (vmbo-tl) to sub-tracks with a more practical focus (vmbo-bbl). The new system also includes tracks for children that necessitate special care, which are vmbo-lwo (special) and vmbo-pro (special+).[6] The perception that the reform made it more difficult to transition from the pre-vocational

[6]The Ministry of Education (2005) reports that about 60% of students attended the pre-vocational track (vmbo) in 2004.

track to higher levels, combined with the fact that the reform brought together pupils from different levels contributed to the image problems of the pre-vocational track, which could have enhanced last place aversion (Ministry of Education, 2005; Kloosterman and de Graaf, 2010).

4.3 Empirical strategy

This paper examines last place aversion in an educational context and uses the reform that caused an exogenous change in the tracks in Dutch secondary education. Tracks that were previously ranked before-last, suddenly became the lowest track. The reform will only affect children who are ex ante expected to attend the new defined lowest track, which will be exploited in a difference-in-difference framework. The difference-in-difference framework allows me to control for factors that are common across survey waves. The outcomes of interest are the variables that determine track allocation in secondary school, namely (1) the track recommendation by the primary school teacher, and (2) the score on the high-stakes cognitive standardized test.[7] Scores on a low-stakes test that is unimportant for track allocation in secondary school are used to identify the group of students that is likely to be affected by the reform.

The difference-in-difference approach compares children who were in 6^{th} grade before the reform with those who were in 6^{th} grade after the reform. The lowest educational track is different for these two groups. Behavioral responses as to avoid the lowest track will only be prevalent among those who are expected to attend the lowest track. This regards efforts made

[7]Information on the actual track allocation in secondary school is not available in the data.

by the child, but similarly parents will only get involved in their child's education if they have the belief that they can improve their outcomes (Hoover-Dempsey and Sandler, 1997). Hence, the treatment group consists of individuals who have a high likelihood of attending the lowest track, and the control group consists of individuals with a low likelihood of attending the lowest track.

To identify which students are more or less likely to attend the lowest track, I use test-scores on a low-stakes cognitive standardized test that is unimportant for track allocation in secondary education. Specifically, I focus on the lowest teacher recommendation that is just above the newly established lowest track. This is the lowest possible track recommendation that would be capable of letting the child avoid attending the lowest track.[8] I then focus on the equivalent track before the reform. The median low-stakes test score for children receiving this pre-reform track recommendation serves as threshold x_i for determining whether children are more likely to attend the lowest track (below x_i) or less likely (above x_i).[9] A necessary assumption for this set-up of treatment and control group is that this low-stakes test score is independent of treatment. This assumption is likely to hold, because (1) the classification is based on the pre-reform distribution, and (2) the low-stakes test is not taken into account for track allocation in secondary education.

Equation 4.1 is estimated for the outcome variables (y_{it}) that determine track allocation in secondary school: the track recommendation, and the

[8]This corresponds to the combination track of the vocational general track and the general track (T6) as is shown in Table 4.3.

[9]An in-depth discussion of the setting up of treatment and control group is provided in Section 4.4. Robustness checks with respect to the chosen threshold are provided in Section 4.5.3 and do not change the conclusions.

high-stakes cognitive test score for child i in year t. The track recommendation variable is categorical with nine possible outcomes, and is transformed to a dummy representing whether the child received a track recommendation above a specified track. I focus on the lowest possible track that will let the child avoid the lowest track.[10] The high-stakes test score is standardized with mean zero and standard deviation one. Equation 4.1 includes an indicator for being in 6^{th} grade after the reform ($Reform_t$), a treatment group indicator stating that a child is more likely to attend the lowest track ($Treatment_i$), and the difference-in-difference coefficient of interest ($Reform_t * Treatment_i$) that represents the effects of the reform for those who are most likely affected by the reform. Vector $\mathbf{X}_i$ includes control variables at the individual level, and ε_{it} is the individual-specific error term, which is clustered at the school-level.

$$\begin{aligned} y_{it} = \gamma_0 + \gamma_1 Reform_t + \gamma_2 Treatment_i + \\ \gamma_3 (Reform_t * Treatment_i) + \mathbf{X}_i \delta + \varepsilon_{it} \end{aligned} \tag{4.1}$$

4.3.1 Validity of the identification strategy

For this approach to be valid a couple assumptions need to hold. First, the policy intervention should be exogenous. Although the changes in the educational system were announced prior to the actual implementation[11], it is impossible for parents and children to manipulate whether or not they are subject to the reform. Parents and children cannot decide that their

[10] The set-up of outcome variables is discussed in more detail in Section 4.4.

[11] Act 25.410 was submitted on 19 June 1997, and accepted by Parliament on 19 May 1998. See: `https://www.eerstekamer.nl/wetsvoorstel/25410_regeling_leerwegen_mavo_en`.

child skips a class to avoid attending the newly introduced lowest vocational track.

Second, there should not be other reforms that were implemented during same time period that could also influence the outcomes of interest. A reform was implemented in 1998 that integrated children with special needs in normal classes. This led to an increase in the number of children in a special needs track from 73,000 in 1998 to 81,000 in 2004 (Ministry of Education, 2005). I perform a robustness check in which I leave out children that are in special needs education, which does not alter my conclusions (see Section 4.5.3). A reform that changed the curriculum of the general (havo) and academic track (vwo) was implemented in 1998/1999. The reform changed the curriculum in the last two (or three for the academic track) years of education, as students had to make a choice for a study field: i.e. science, health, social sciences or humanities. This reform could have affected the perceived transition probabilities from the vocational track to these higher tracks, which could have contributed to an increased level of last place aversion.

Third, the dataset used in this paper is a repeated cross-section, implying that the distributions of ability must be similar before and after the reform. Hence, a different composition of students should not be driving the results found. This is particularly important as the composition of included schools changes by survey wave. Figure 4.1 shows the distributions of the low-stakes PRIMA test before and after the reform.[12] As this test is not important for track allocation it is likely unaffected by the reform. The distributions for reading and math look very similar before and after the

[12]The distributions by survey wave are shown in Figure 4.A1.

reform, although a Kolgomorov-Smirnov test rejects the null of equality of distributions. To take into account that the composition of schools changes over the waves, I perform a robustness check in which I only include schools that are present in all waves which does not change my conclusions (see Section 4.5.3).

Finally, the key assumption for a difference-in-difference strategy is that the trends in outcomes would have been similar in the treatment and control group in absence of treatment, which is also known as the common trend assumption. Unfortunately I can not go more than three years back due to data availability. Figure 4.2 shows the common trend plots for the two outcomes of interest for the treatment groups based on the reading and math score for the available data. The pre-trends in outcomes before the reform look similar in treatment and control group.

4.4 Data

The data on cognitive test scores and the teacher assessments used in this paper comes from the longitudinal PRIMA survey. The survey started to learn more about the ability of primary school children in the Netherlands (Kamphuis et al., nd). The survey was administered biannually from 1994 to 2005, and data from all six waves are used (1994-1995, 1996-1997, 1998-1999, 2000-2001, 2002-2003, and 2004-2005). This implies that I have three waves of data before the reform and three waves of data after the reform. Each wave of survey data contains information on about 60,000 students from about 600 schools.[13] The survey is completed for students in year 2, 4,

[13]The sample composition changed per wave, but Roeleveld and Vierke (2003) find no significant differences between schools that exited the PRIMA-project and schools that did not drop out.

6, and 8 of Dutch primary education, which is equivalent to grade 0, grade 2, grade 4 and grade 6. The PRIMA survey contains information on the standardized cognitive Cito-test, the teacher recommendation of the child, information on the socioeconomic background of the child, and finally also the child's performance on a standardized test unrelated to track allocation in secondary school.

The outcomes of interest are the variables that determine track allocation in secondary school, i.e. the child's performance on the standardized cognitive test taken in the final year of primary education (Cito-test), and the track recommendation made by the child's teacher. Note that I only observe the variables that *determine* track allocation in secondary school, but unfortunately cannot observe the *actual* track assignment in secondary school. Scores for the standardized cognitive Cito-test range from 501-550, and are standardized by wave with mean zero and a standard deviation of one.

The teacher's track recommendation variable can take on multiple values. Teachers can give track recommendations for one track (e.g. the academic track (vwo)), but can also specify combined recommendations (e.g. a combination between the academic track (vwo) and the general track (havo)). The possible outcomes of the teacher track recommendation before and after the reform are documented in Table 4.2. There are eleven possible track recommendation outcomes before the reform and fifteen potential outcomes after the reform. A classification is made to map the track recommendations before the reform into track recommendations after the reform. I use the mapping as suggested by Timmermans et al. (2013), who match each track to the amount of years a child is expected to

attend education based on his or her track recommendation. The resulting classification in shown in Table 4.3 and has nine possible track recommendations. The track recommendation variable is coded according to this classification with T1 representing the lowest track recommendation, and T9 representing the highest track recommendation.

The PRIMA survey also contains information on a low-stakes standardized test that is not important for track allocation in secondary school: the PRIMA-test. It is a cognitive test that focuses on reading and math skills, and it is identical for different grades in different years (Chorny et al., 2010). The PRIMA-test has an ordinal score (Leuven et al., 2010), and scores are standardized by wave and by grade with mean zero and a standard deviation equal to one.

This low-stakes standardized test is used to define a treatment and control group. The idea is that children are only affected by the reform, and hence last place aversion, if they are expected to attend the lowest track. Hence there could be behavioral responses that affect the variables that determine track allocation such that the newly developed lowest track is avoided. When referring back to Table 4.3 the lowest track recommendation that would make a child avoid the lowest track is the combination advice of the vocational general and the general track (T6). I then examine the average low-stakes PRIMA for children receiving a T6 advice before the reform. This is an indicator for the average ability level that would correspond to a track recommendation just above the lowest track, and hence would be an indicator of which children are more likely to display behavioral responses after the reform. Table 4.4 shows the different tracks and the corresponding mean scores on the low-stakes cognitive PRI-

MA test. Higher track recommendations are associated with higher mean and median scores on both math and reading. I use the median scores to set up two treatment groups, one based on the math score and one based on the reading score. This implies that a child is assigned to the treatment group when having a PRIMA reading score below 0.165, and a PRIMA math score below 0.234.[14]

The dataset also includes information on gender and age (in months), and information on the socioeconomic background of the child. The child's socioeconomic background is proxied with the subsidy-factor, which represents the amount of (extra) funding schools receive for each student. Schools receive higher subsidies for children from lower socioeconomic strata. Three main groups can be distinguished: (i) non-disadvantaged students have a subsidy factor equal to 1, (ii) native students with lower educated parents have a subsidy factor of 1.25, (iii) the subsidy factor for students from ethnic minorities is equal to 1.9.[15] The value of the subsidy-factor represents the amount of extra funding schools receive for each student, i.e. schools receive 25% extra funding for students whose subsidy-factor is equal to 1.25 (Chorny et al., 2010). I create a dummy which is equal to zero if the child is not disadvantaged (subsidy-factor equal to one) and one if the child is disadvantaged (subsidy-factor higher than one). The data also includes information on the educational level of the parents.

This paper focuses on variables determining the transition from primary school to secondary school and therefore only uses data from students

[14]The results are robust to using different definitions of the treatment group (Section 4.5.3).

[15]There are groups of pupils with subsidy factors equal to 1.4 and 1.7, but these categories each only account for less than 1% of the sample.

in the 6^{th} grade. The data for the sample of students in only the 6^{th} grade contains information on $84,895$ individuals from $1,592$ schools. Children with missings on key demographic variables, i.e. with missing age (N = $4,061$), missing gender (N = $7,238$), and missing information on socio-economic background (N = $6,185$) are dropped from the sample. In the remaining sample of $74,664$ children, I observe the high-stakes standardized cognitive test scores for $47,713$ children, the teacher recommendation for $59,981$ children. The lower number of observations for the Cito-test is not surprising as administering the test is not mandatory for schools. The low-stakes standardized cognitive test for math and reading performance is observed for $68,554$ and $71,011$ children respectively.

This paper focuses on the variables that determine track allocation in secondary school. For the estimation sample I drop children for whom either the track recommendation, the Cito-score, or the low-stakes standardized test scores for math and reading are missing (N = $32,286$). This high number is mainly caused by a high number of missing for the Cito-score (N = $26,951$). An overview of the differences between the full sample and the estimation sample is documented in Table 4.A2. Children in the estimation sample have higher test scores, but are very similar in background characteristics as compared to the full sample.

Descriptive statistics for estimation sample by treatment assignment are shown in Table 4.5. Children in the treatment group have significantly lower test-scores as compared to those in the control group, and they are significantly less likely to have a track recommendation above T6. Children in the treatment group are also significantly older, which suggests that grade retention is more prevalent in this group. Children in the treatment

group are significantly more disadvantaged and the share of children that has at least one parent with higher education is significantly lower.

4.4.1 Track assignment variables pre- and post-reform

The reform changed the lowest track, and if last place aversion plays a role in this educational context, the expectation is that the variables that determine track allocation in secondary school are affected by the reform. The distribution of track recommendations before and after the reform is shown in Figure 4.3 (the classification of track recommendations refers to Table 4.3). The lowest track recommendation that would avoid the lowest track is T6 (the combination advice to the vocational track and the general track). The figures show that there is a shift in the type of track recommendations children receive (note that the figure does not control for the underlying ability of children).[16]

More children receive a T5 track recommendation, and less children receive track recommendations T2 and T4. It is not the case that more children get a T6 recommendation, and hence receive a track recommendation that would allow them to just escape the lowest track. Another change occurs in T1, which is the track recommendation for children in special education. The amount of children receiving a recommendation in this category almost doubled, which can be explained by a policy reform implemented in 1998 that integrated children with learning- and behavioral difficulties into the vocational track.

The distribution of the high-stakes standardized Cito-test before and after the reform is shown in Figure 4.4. The pre-reform distribution looks

[16]The Kolgomorov-Smirnov test rejects the null hypothesis of equality of the track recommendation distributions before and after the reform (at the 1% level).

rather smooth, whereas the post-reform distribution shows some irregularities.[17] Note that these irregularities cannot be explained by a manipulation of test-scores by teachers, as test sores are centrally determined by the Centraal Instituut voor Toetsontwikkeling (CITO) after the test is sent in by the teacher. Three main irregularities are visible in the figure. The first is visible after a score of 524, which is the score necessary for qualifying for a vocational lower track recommendation (vmbo-kb).[18] The second is visible after a score of 529, which is the score corresponding to a vocational general track recommendation (vmbo-tl). Finally, a bump is visible after a score of 537 which is the score corresponding to a general track recommendation (havo). This suggests that there may have been investments in extra tutoring to get the child the score that matches with the desired track.

4.5 Results

4.5.1 OLS estimates

Table 4.6 shows the results from OLS models regressing the outcome variable on a dummy representing the introduction of the reform. The dummy is insignificant in the specification without controls for both outcome variables (column 1). Column 2 adds controls for the child's background characteristics (i.e. age, gender, and an indicator for being disadvantaged), after which the reform-dummy becomes significant for both outcome variables. Conditional on demographic characteristics, children have a 2.9 percentage

[17] The Kolgomorov-Smirnov test does not reject the null of equality of Cito-score distributions.

[18] The thresholds connecting the Cito-score to the track types are shown in Table 4.A1.

point lower probability of attending a track above the lowest track, and children have a lower Cito-score by 7% of a standard deviation after the reform.

However, when adding controls for the ability of the child, i.e. their performance on the low-stakes standardized test, the sign of the coefficients changes. Conditional on background characteristics and ability, children have a 1.2 percentage points higher chance of attending a track above the lowest track after the reform, and a higher Cito-score by 3.5% of a standard deviation. It could be that wave-specific effects, e.g. different schools included in the sample, are driving the effects found in a simple before-after comparison. I employ a difference-in-difference strategy to control for these wave-specific factors.

4.5.2 Difference-in-difference estimates

The difference-in-difference approach compares children who took the test before the reform with those who took the test after the reform, and uses the fact that some children are more likely to be affected by the reform (i.e. the treatment group).

The results of this approach for track recommendation as the relevant outcome are shown in Table 4.7. The outcome variable is a dummy representing that the track recommendation is equal or higher than T6. The child would be able to just avoid the lowest track with this track recommendation. Column 1 and 2 use the treatment definition based on the PRIMA reading score, whereas column 3 and 4 use the treatment definition based on the math PRIMA score. The coefficient of interest ($Treatment * Reform$) is very similar in the specification with and without controls. Children who

are likely to attend the lowest track based on their reading score have a 1.9 percentage points lower probability of receiving a track recommendation above the lowest track after the reform. The effect is more negative in the treatment specification based on the math score: children who are likely to attend the lowest track have a 6.5 percentage points lower probability of receiving a track recommendation above the lowest track. This effect is large compared to a mean of 44.5%.

Table 4.8 shows the results with the high-stakes cognitive Cito-score is the relevant outcome variable. The coefficient of interest is insignificant when treatment assignment is based on the reading score. When treatment is based on the math score, children in the treatment group have a significantly lower cognitive test-score after the reform (5% of a standard deviation). This result is consistent with the finding for the track recommendation. The Cito-score matters for the teacher's track recommendation, and hence track recommendations could be lower if test-scores are lower. I examine the robustness of these results and possible mechanisms that can explain these results in the following sections.

4.5.3 Robustness checks

The difference-in-difference estimates show that children who are likely to attend to lowest track have lower test scores after the reform, and are less likely to have a track recommendation above the lowest track. This section discusses robustness checks with respect to the chosen identification strategy.

First, I check the robustness of the results to the chosen treatment group. The treatment group is defined based on the median pre-reform

PRIMA-score of the distribution that corresponds to a track recommendation above the lowest track. Table 4.9 shows the robustness of this definition to the mean score, the 25th, and 75th percentile scores. The point-estimates are very similar as compared to the baseline specification. I also perform a robustness check in which I use the baseline specification of the treatment group, but leave out children with PRIMA-scores between the 25th and 75th percentile. Children who are on the border between being less or more likely to attend the lowest track are dropped, and this thus creates a more clean treatment and control group. This practice yields similar results, and if anything the point-estimates are larger (more negative for both outcome variables).

Second, as the composition of schools changes over the survey waves that are included, I restrict my sample to schools that are present in all six waves. Table 4.9 shows that the results based on the reading treatment group are very similar. Those for the math treatment group increase in magnitude. To illustrate, children in the treatment group have a 11.9 percentage points lower probability of receiving a track recommendation above the lowest track, and they have a lower test score by 8.5% of a standard deviation. Point estimates are very similar after adding school fixed effects as well.

Third, I perform a robustness check in which I leave out children who received a track recommendation for a special or special+ track. Another reform was implemented in the year prior to the reform that integrated these students in regular classes. Leaving out these students yields fairly similar results. Thus, the results are not sensitive to the inclusion of these special care students.

Finally I also execute a placebo check in which I impose that the introduction of the reform occurred between survey waves two and three, instead of the actual introduction between survey wave three and four. The results in Table 4.9 show that the coefficient of interest is significantly different from zero for the Cito-score and for the track recommendation (when treatment is based on the math score). Hence, there are also responses for those most likely to attend the lowest track a survey wave prior to the reform. This could have to do with another reform was introduced in the academic year of 1998 which changed the curriculum of the highest academic tracks. The perceived ease of transitioning from the lowest educational track to these higher tracks could have been affected by this reform, which may explain the significant point-estimates for this placebo check.

4.6 Mechanisms at work

Children who are likely to attend the lowest track (treatment group) are less likely to avoid the lowest track, and have lower test-scores after the reform. This section explores several explanations for this finding.

First, it is interesting to see what happens to the child's track recommendations, given that children in the treatment group are less likely to receive a track recommendation of T6 or higher. I estimate the difference-in-difference models for dummy outcome variables that indicate that a child received a track recommendation equal or above T3, T4, T5, T7, and T8. Table 4.10 shows that children who are expected to attend the lowest track are more likely to receive a T5 or higher track recommendation after the reform, hence it seems that the reform does not lead to increased track

recommendations above the lowest track, but rather to more recommendations for the highest sub-track of the new lowest track.

Second, I examine whether lower test scores can explain the lower track recommendations. If last place aversion causes parents and children to exert extra effort in order to avoid the lowest track, they have a couple of possibilities at their disposal. One way to avoid the lowest track is by performing very well on the high-stakes standardized cognitive test. Investments in homework assistance, or higher test motivation, could lead to higher scores. Parents can also influence the track recommendation by putting pressure on the primary school teacher in order for them to give a higher recommendation. It is very hard to separate the two channels with the available data. I add a control for the high-stakes standardized test-score to the model with the track recommendation as outcome variable. The coefficient of interest is hypothesized to be no longer significant with the inclusion of this control if only the high-stakes test score matters for the track recommendation. Table 4.11 shows that the point estimates are still negative and significant when the Cito-score is added as an extra control. This suggests that teachers give a lower track recommendation even conditional on the child's performance on the high-stakes test.

Third, I examine heterogeneity by subject matter. The Cito-score is an aggregate score, and it measures performance in reading, math, and information processing. The treatment groups are defined based on math and reading performance on a low-stakes cognitive standardized, and hence it is interesting to see whether and how the results change when a distinction is made between math and reading performance in the outcome variable. This is particularly interesting as Table 4.11 shows that the point-estimates

are somewhat different when estimating the models for a sample of girls and boys. For example, the effects for girls on both outcome variables are stronger when treatment is determined by the math score. This could potentially be explained by gender differences in educational performance (e.g Fryer and Levitt, 2010).

The Cito-scores for math and reading performance are not available in all survey waves, which explains the lower number of observations. Scores are standardized by wave with mean zero and standard deviation of one. Table 4.12 shows the results of estimating the models for these new outcome variables. The first row pools the sample of boys and girls and shows that the point-estimate for reading is only negative and significant for those who are in the treatment group based on their reading score. Similarly the point estimate for math is only significantly negative for those who are in the treatment group based on their math score. When separating by gender, it is interesting to see that the negative effect on reading performance (based on the reading treatment) is primarily driven by boys. Whereas the negative effect on math performance (based on the math treatment) is stronger for girls. The child's weakest subject may be the one where pressure to perform well is higher. Hence, these results suggest that children do particularly worse after the reform in their weakest subject, the one where the pressure to perform well is hypothesized to be higher.

Finally, last place aversion could be stronger among children born in families with highly educated parents. Higher educated parents have a higher valuation of education (Hastings et al., 2005; Burgess et al., 2015), and may be more aware of differences in educational quality (Sacerdote et al., 2011; Borghans et al., 2015). Kloosterman and de Graaf (2010) show

that in the Netherlands, higher educated parents prefer grade retention for their child instead of moving one track down. I examine heterogeneity by parental education level. I define parents to be highly educated if at least one of the two attended either higher vocational or university education.[19] Table 4.11 shows that the point-estimates are more negative for children with higher-educated parents. This is suggestive evidence for the fact that children are more negatively affected by the reform if the pressure to perform well is higher.

4.7 Conclusion

Last place aversion describes a situation in which individuals dislike being in last place. This paper studies how last place aversion affects the child's educational outcomes, and specifically outcomes that are important for ability tracking in secondary education. I exploit a Dutch educational reform that merged the two lowest tracks in secondary education. Hereby a track that was previously not the lowest track, suddenly became the lowest track. I combine the timing of the reform with the fact that only children who are expected to attend the lowest track are affected by the reform. The children who are expected to attend the lowest track are identified by their performance on a standardized cognitive test that is unimportant for track allocation. The outcomes of interest are the variables that determine track allocation in secondary school: the child's performance on a high-stakes standardized cognitive test, and the child's track recommendation provided by the primary school teacher.

[19]Note that the number of observations is lower as information on parental education is not available for all children.

I find that children who are expected to attend the lowest track are less likely to receive a track recommendation above the lowest track, and perform worse on the high-stakes standardized test. The lower track recommendations cannot fully be explained by lower test performance. If this reform represents a good setting to study last place aversion, these findings suggest that last place aversion leads to lower educational outcomes. Furthermore, the effects are concentrated among the child's weakest subject matter, which may be the subject for which the pressure to perform well is higher. I also find that the effects are stronger in higher-educated families: families where last place aversion is hypothesized to be stronger. This suggests that last place aversion leads to worse performance, as the pressure to perform well is higher.

There are a few limitations of this paper. Unfortunately the data only contains information on the variables that determine track allocation but not on actual track allocation in secondary school. Also, I cannot observe outcomes during secondary school (e.g. track switching) or outcomes after secondary school (e.g. wages). Moreover, I can only provide suggestive evidence that the effects are stronger when the pressure to perform is higher. With the current data it is impossible to observe parental investment/pressure, and particularly leading up to the examinations in the final year of primary education. Future research should address how last place aversion shapes parental investments, how it affects actual track allocation in secondary school, and outcomes later in life.

Tables

Table 4.1: Pre- and Post-reform Dutch secondary school tracks

	Pre-reform (1)	Post-reform (2)	
Academic track	VWO	VWO	
General track	HAVO	HAVO	
Vocational track	MAVO	VMBO	Theoretical (*vmbo-tl*)
			Mixed (*vmbo-gl*)
	LBO		Middle (*vmbo-kbl*)
			Basic (*vmbo-bbl*)
Special needs education			
Special track	Special (IVBO)		Special (*vmbo-lwo*)
	Special+ (VSO)		Special+ (*vmbo-pro*)

Table 4.2: Track recommendation outcomes, pre- and post-reform

Pre-reform		Post-reform	
Academic	vwo	Academic	vwo
General/academic	havo/vwo	General/academic	havo/vwo
General	havo	General	havo
General/general vocational	mavo/havo	Vocational theor./general	vmbo-tl/havo
General vocational	mavo	Vocational theoretical	vmbo-tl
Lower/general vocational	vbo/mavo	Vocational mixed/theoretical	vmbo-gl/tl
Lower vocational	vbo	Vocational mixed	vmbo-gl
Vocational special	ivbo/vbo	Vocational middle/mixed	vmbo-kbl/gl
Vocational special	ivbo	Vocational middle	vmbo-kbl
Vocational special+/special	vso/ivbo	Vocational basic/middle	vmbo-bbl/kbl
Vocational special+	vso	Vocational basic	vmbo-bbl
		Vocational special/basic	vmbo-lwo/bbl
		Vocational special	vmbo-lwo
		Vocational special/special+	vmbo-pro/lwo
		Vocational special+	vmbo-pro

Table 4.3: Track recommendation, pre- and post-reform mapping

		Pre-reform	Post-reform
T9	Academic	vwo	vwo
T8	Academic/general	havo/vwo	havo/vwo
T7	General	havo	havo
T6	Vocational general/general	mavo/havo	vmbo-gl/-tl/havo
T5	Vocational general	mavo	vmbo-gl/tl
T4	Vocational lower/general	vbo/mavo	vmbo-bbl/kbl/gl/tl
T3	Vocational lower	vbo	vmbo-bbl/kbl
T2	Vocational special	ivbo/vbo	vmbo-lwo/bbl/kbl
T1	Vocational special/special+	vso/ivbo	vmbo-pro/lwo

Notes: Classification based on Timmermans et al. (2013).

Table 4.4: Treatment and control group classification

Track	PRIMA Reading			PRIMA Math		
	Mean	Median	N	Mean	Median	N
T9	1.291	1.252	1,914	1.341	1.278	1,886
T8	0.872	0.791	2,906	0.898	0.862	2,873
T7	0.545	0.496	3,455	0.600	0.546	3,402
T6	0.222	0.165	3,445	0.243	0.234	3,413
T5	-0.061	-0.058	5,028	-0.064	-0.086	4,977
T4	-0.320	-0.329	2,878	-0.376	-0.402	2,846
T3	-0.596	-0.594	5,619	-0.665	-0.691	5,502
T2	-0.966	-0.961	411	-1.137	-0.139	389
T1	-1.107	-1.137	1,100	-1.291	-1.334	995

Notes: The full sample pre-reform scores are reported.

Table 4.5: Descriptive statistics: treatment versus control

	Reading based		Math based			
	Treatment	Control	Treatment	Control		
	(1)	(2)	(3)	(4)	(1)-(2)	(3)-(4)
	N=23,381	N=18,997	N=23,755	N=18,623		
Track $\geq$ T6 (0/1)	0.225	0.715	0.187	0.773	***	***
	(0.418)	(0.452)	(0.390)	(0.419)		
Cito test-score	-0.450	0.687	-0.512	0.789	***	***
	(0.870)	(0.690)	(0.824)	(0.591)		
PRIMA reading score	-0.617	0.933	-0.316	0.580	***	***
	(0.549)	(0.699)	(0.832)	(0.946)		
PRIMA math score	-0.323	0.566	-0.605	0.943	***	***
	(0.871)	(0.888)	(0.626)	(0.597)		
Age (average)	12.183	11.967	12.176	11.971	***	***
Female (0/1)	0.498	0.500	0.553	0.431		***
Disadvantaged (0/1)	0.638	0.330	0.607	0.363	***	***
One parent higher educated (0/1)	0.077	0.211	0.081	0.209	***	***

Notes: Standard deviations in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Based on the estimation sample.

Table 4.6: OLS estimates

	I(T≥T6)		
	(1)	(2)	(3)
Introduction vmbo	-0.003	-0.029***	0.012*
	(0.009)	(0.007)	(0.007)
Demographic controls	N	Y	Y
Ability controls	N	N	Y
R^2	0.000	0.133	0.450
N	42,378	42,378	42,378
	Cito-score		
	(1)	(2)	(3)
Introduction vmbo	-0.009	-0.070***	0.035***
	(0.023)	(0.018)	(0.011)
Demographic controls	N	Y	Y
Ability controls	N	N	Y
R^2	0.000	0.184	0.726
N	42,378	42,378	42,378

Notes: Estimated by OLS, estimation sample. I($T \geq T6$) represents a dummy variable which is equal to one when the child received a track recommendation greater or equal than the combination advice of the vocational general and the general track (T6). Demographic controls include gender, age (linear and quadratic) and an indicator for being disadvantaged. The ability controls include the low-stakes PRIMA test-score for math and reading. Standard errors are within brackets and are clustered at the school level.

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

Table 4.7: Difference-in-difference results: track recommendation

	I(T≥T6)			
	(1)	(2)	(3)	(4)
Reform	-0.003	0.023***	0.018**	0.043***
	(0.009)	(0.009)	(0.009)	(0.009)
Treatment	-0.406***	-0.128***	-0.502***	-0.217***
	(0.007)	(0.008)	(0.007)	(0.011)
Treatment*Reform	-0.014	-0.019**	-0.057***	-0.065***
	(0.009)	(0.008)	(0.009)	(0.009)
Demographic controls	Y	Y	Y	Y
Ability controls	N	Y	N	Y
Treatment	Reading	Reading	Math	Math
Mean dep. var.	0.445	0.445	0.445	0.445
R^2	0.284	0.458	0.390	0.476
N	42,378	42,378	42,378	42,378

Notes: Estimated by OLS, estimation sample. I($T \geq T6$) represents a dummy variable which is equal to one when the child received a track recommendation greater or equal than the combination advice of the vocational general and the general track (T6). Demographic controls include gender, age (linear and quadratic) and an indicator for being disadvantaged. The ability controls include the low-stakes PRIMA test-score for math and reading. Standard errors are within brackets and are clustered at the school level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.8: Difference-in-Difference: Cito-score

	Cito-score (std)			
	(1)	(2)	(3)	(4)
Reform	-0.044***	0.026**	-0.026**	0.059***
	(0.014)	(0.012)	(0.013)	(0.011)
Treatment	-0.982***	-0.150***	-1.158***	-0.154***
	(0.015)	(0.012)	(0.015)	(0.026)
Treatment*Reform	0.032*	0.017	-0.021	-0.050***
	(0.019)	(0.013)	(0.019)	(0.016)
Demographic controls	Y	Y	Y	Y
Ability controls	N	Y	N	Y
Treatment	Reading	Reading	Math	Math
Mean dep. var.	0.060	0.060	0.060	0.060
R^2	0.397	0.728	0.503	0.729
N	42,378	42,378	42,378	42,378

Notes: Estimated by OLS, estimation sample. Demographic controls include gender, age (linear and quadratic) and an indicator for being disadvantaged. The ability controls include the low-stakes PRIMA test-score for math and reading. Standard errors are within brackets and are clustered at the school level.

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

Table 4.9: Robustness results

	I(T≥T6)		Cito-score (std)	
	Reading	Math	Reading	Math
	(1)	(2)	(3)	(4)
Baseline	-0.019**	-0.065***	0.017	-0.050***
	(0.008)	(0.009)	(0.013)	(0.016)
N	42,378	42,378	42,378	42,378
Treatment - mean	-0.019**	-0.065***	0.015	-0.058***
	(0.009)	(0.009)	(0.013)	(0.016)
N	42,378	42,378	42,378	42,378
Treatment - 25th perc.	-0.017*	-0.055***	0.037***	-0.032*
	(0.009)	(0.010)	(0.014)	(0.018)
N	42,378	42,378	42,378	42,378
Treatment - 75th perc.	-0.001	-0.061***	0.027**	-0.086***
	(0.009)	(0.009)	(0.014)	(0.016)
N	42,378	42,378	42,378	42,378
Treatment - excl. 25-75th perc.	-0.010	-0.077***	0.044***	-0.075***
	(0.010)	(0.010)	(0.016)	(0.020)
N	27,193	27,696	27,193	27,696
Schools all waves	-0.020	-0.119***	0.033	-0.085***
	(0.019)	(0.018)	(0.029)	(0.029)
N	9,066	9,066	9,066	9,066
Schools all waves (fixed effects)	-0.010	-0.104***	0.033	-0.091***
	(0.017)	(0.017)	(0.027)	(0.028)
N	9,066	9,066	9,066	9,066
Without special(+)	-0.025***	-0.071***	0.036***	-0.029*
	(0.009)	(0.010)	(0.013)	(0.016)
N	39,482	39,482	39,482	39,482
Placebo (reform at T=2)	-0.005	-0.090***	0.061***	-0.078***
	(0.009)	(0.010)	(0.015)	(0.017)
N	42,378	42,378	42,378	42,378

Notes: Double-difference estimate $Treatment * Reform$ is reported, full control specification.

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

Table 4.10: Other track recommendations

	Reading	Math
	(1)	(2)
Outcome: I(T≥T3)	-0.041***	-0.049***
	(0.006)	(0.006)
N	42,378	42,378
Outcome: I(T≥T4)	-0.036***	-0.063***
	(0.009)	(0.010)
N	42,378	42,378
Outcome: I(T≥T5)	0.033***	0.026***
	(0.008)	(0.010)
N	42,378	42,378
Outcome: I(T≥T6)	-0.019**	-0.065***
	(0.008)	(0.009)
N	42,378	42,378
Outcome: I(T≥T7)	-0.032***	-0.080***
	(0.009)	(0.010)
N	42,378	42,378
Outcome: I(T≥T8)	-0.038***	-0.062***
	(0.010)	(0.011)
N	42,378	42,378

Notes: Double-difference estimate $Treatment * Reform$ is reported, full control specification.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.11: Mechanisms at work

	I(T≥T6)		Cito-score (std)	
	Reading	Math	Reading	Math
	(1)	(2)	(3)	(4)
Baseline	-0.019**	-0.065***	0.017	-0.050***
	(0.008)	(0.009)	(0.013)	(0.016)
N	42,378	42,378	42,378	42,378
+Cito control	-0.025***	-0.051***		
	(0.008)	(0.008)		
N	42,378	42,378		
Boys	-0.018	-0.052***	-0.000	-0.047**
	(0.011)	(0.012)	(0.016)	(0.020)
N	21,218	21,218	21,218	21,218
Girls	-0.020*	-0.083***	0.037**	-0.058***
	(0.011)	(0.011)	(0.017)	(0.018)
N	21,160	21,160	21,160	21,160
Lower-educated	-0.016*	-0.060***	0.016	-0.061***
	(0.010)	(0.010)	(0.014)	(0.016)
N	33,858	33,858	33,858	33,858
Higher-educated	-0.021	-0.083***	-0.029	-0.080**
	(0.025)	(0.027)	(0.033)	(0.038)
N	5,373	5,373	5,373	5,373

Notes: Double-difference estimate $Treatment * Reform$ is reported, full control specification.

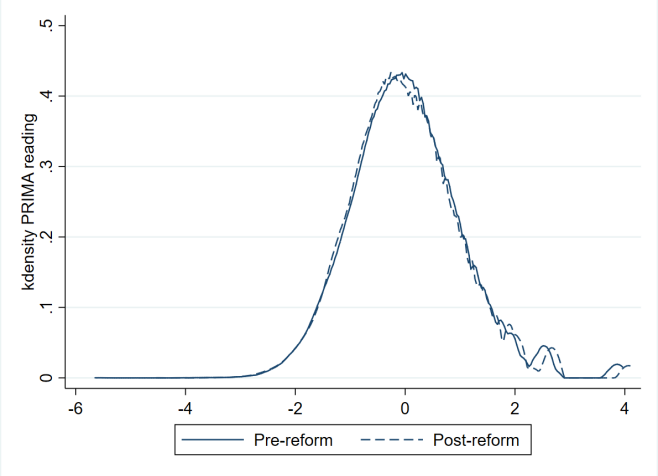
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.12: Mechanisms at work

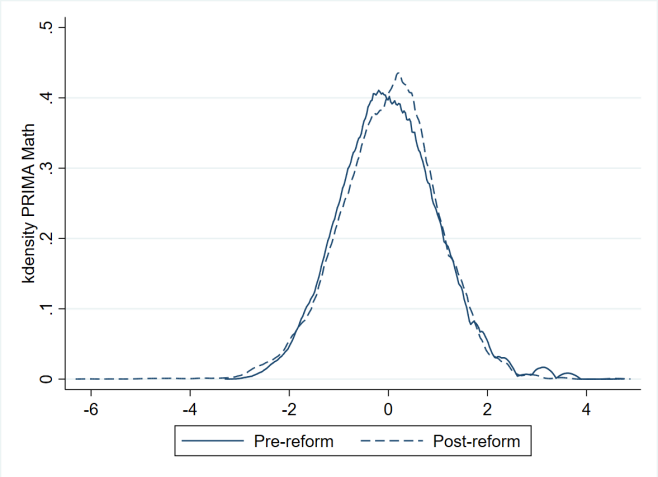
	Cito-Reading (std)		Cito-Math (std)	
	Reading	Math	Reading	Math
	(1)	(2)	(3)	(4)
Boys and girls	-0.038**	0.016	0.031*	-0.054***
	(0.019)	(0.020)	(0.017)	(0.020)
N	35,545	35,545	35,538	35,538
Boys	-0.067***	0.034	0.019	-0.050**
	(0.025)	(0.026)	(0.021)	(0.025)
N	17,889	17,889	17,873	17,873
Girls	-0.006	0.008	0.043**	-0.057***
	(0.023)	(0.024)	(0.020)	(0.021)
N	17,656	17,656	17,665	17,665

Notes: Double-difference estimate $Treatment * Reform$ is reported, full control specification.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figures

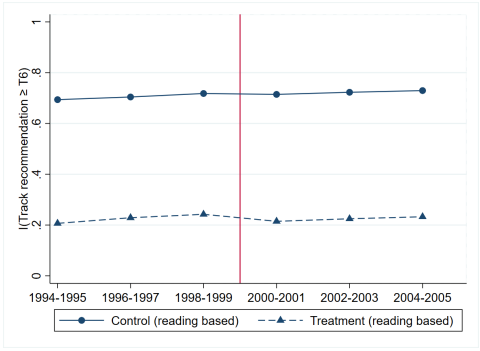


(a) PRIMA Reading

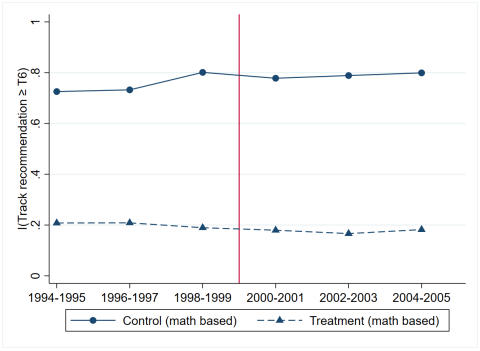


(b) PRIMA Math

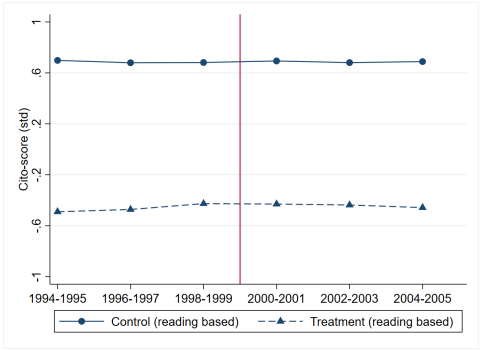
Figure 4.1: Kernel densities before and after the reform, full sample.



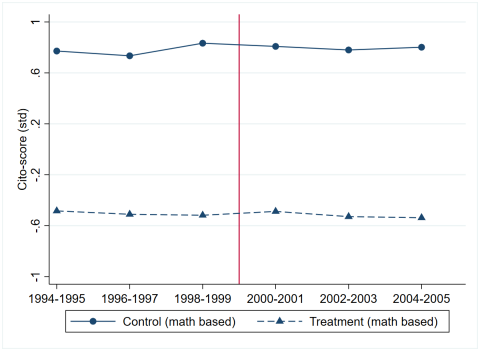
(a) Track ≥ T6 (0/1)



(b) Track ≥ T6 (0/1)

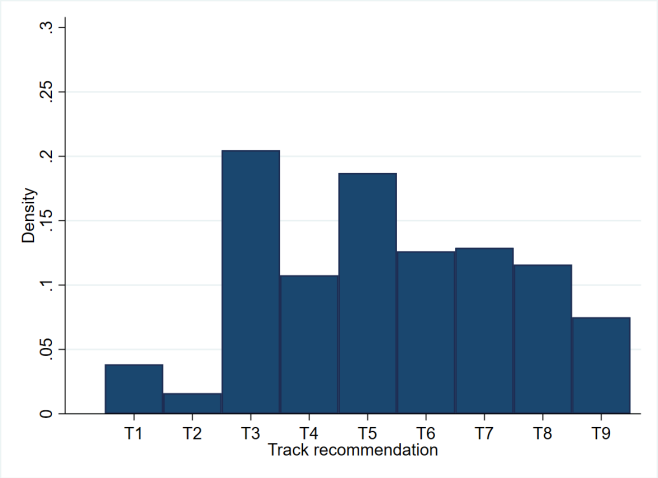


(c) Cito-score (std)

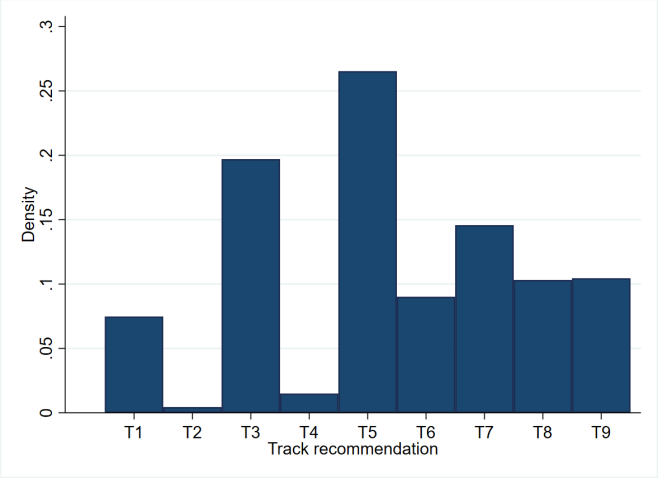


(d) Cito-score (std)

Figure 4.2: Common trend in outcomes, estimation sample.



(a) Pre-reform



(b) Post-reform

Figure 4.3: Track recommendation, pre- and post-reform, estimation sample.

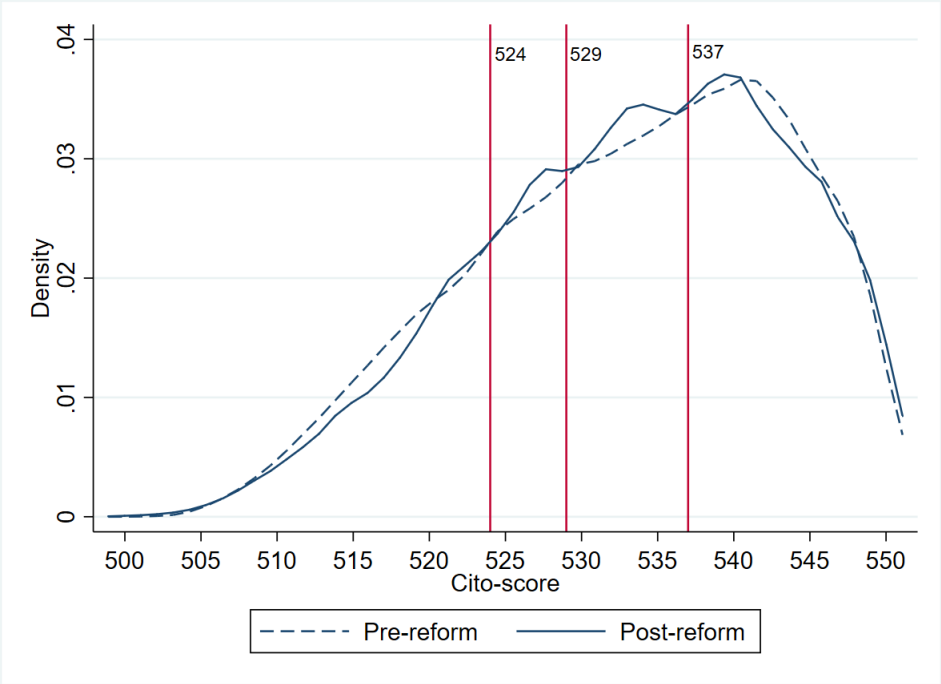


Figure 4.4: Kernel density of Cito-scores pre- and post-reform

Appendix

Table 4.A1: Cito-score and corresponding school tracks.

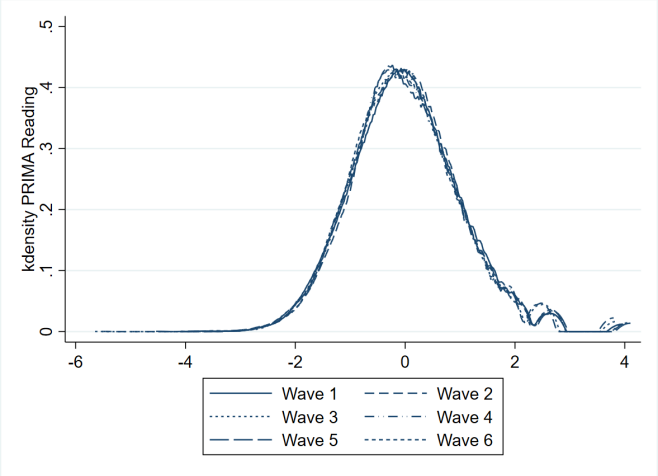
Score range	Track
501-523	vmbo-bb
524-528	vmbo-kb
529-536	vmbo-gl/vmbo-tl
537-544	havo
545-550	vwo

[1] Classification by Van Boxtel et al. (2011).

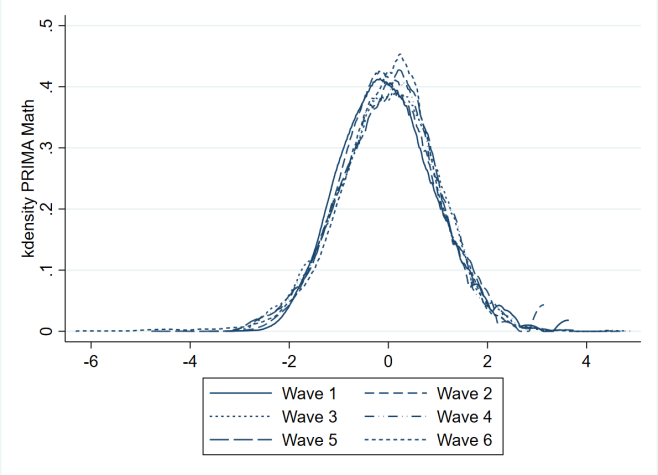
Table 4.A2: Sample selection overview

	Full sample (1)		Estimation sample (2)	
	Pupils	*Schools*	*Pupils*	*Schools*
1994-1995	11,743	651	6,120	379
1996-1997	10,512	603	5,930	379
1998-1999	12,126	592	6,507	378
2000-2001	13,276	588	7,776	422
2002-2003	13,607	589	8,194	420
2004-2005	13,400	587	7,851	406
N	74,664	1,475	42,378	1,061
Cito test-score	0.016		0.060	
	(0.997)		(0.975)	
Participation (%)	0.64		1.00	
PRIMA math score	0.011		0.076	
	(0.996)		(0.983)	
PRIMA reading score	0.013		0.078	
	(0.100)		(0.989)	
Track recommendation	N	%	N	%
T1	4,286	5.74	2,494	5.89
T2	578	0.77	402	0.95
T3	12,280	16.45	8,492	20.04
T4	3,425	4.59	2,354	5.55
T5	13,612	18.23	9,792	23.11
T6	6,323	8.47	4,490	10.60
T7	8,165	10.94	5,864	13.84
T8	6,106	8.18	4,608	10.87
T9	5,206	6.97	3,882	9.16
Missing	14,683	19.67	0	0.00
Age (average)	12.10		12.09	
Gender (% girl)	50.2		49.9	
One parent higher educated (0/1)	0.136		0.137	
Subsidy factor (%)				
1.00	49.48		49.99	
1.25	26.02		27.05	
1.40	0.27		0.28	
1.75	0.37		0.37	
1.90	23.85		22.31	

Notes: Standard deviations in brackets. The number of observations is lower for the parental education variable, N= $67,745$ for the full sample and N = $39,231$ for the estimation sample.



(a) PRIMA Reading



(b) PRIMA Math

Figure 4.A1: Kernel densities by survey wave, full sample.

Nederlandse samenvatting (Summary in Dutch)

In dit proefschrift onderzoek ik hoe omstandigheden in de kinderjaren de vorming van menselijk kapitaal beïnvloeden. Het proefschrift bevat drie hoofdstukken die elk een van de factoren bestuderen die van invloed kunnen zijn op menselijk kapitaal. De focus van dit proefschrift ligt op de rol van de familie - van voor de conceptie van het kind tot lang na de geboorte.

Hoofdstuk twee analyseert hoe betere sociaaleconomische omstandigheden ouderlijke selectie beïnvloeden (welke ouders een kind krijgen), en uiteindelijk hoe dit doorwerkt in de arbeidsmarkt- en gezondheidsuitkomsten van het kind. Ik gebruik hiervoor regionale variatie die voortkomt uit het einde van de Tweede Wereldoorlog in Nederland. In dit hoofdstuk vind ik geen bewijs dat betere sociaaleconomische omstandigheden leiden tot ouderlijke selectie. Verder vind ik voor een groep ongeplande concepties, die zijn voortgekomen uit de betere sociaaleconomische omstandigheden, dat opgroeien in een minder stabiele familie niet leidt tot slechtere arbeidsmarkt- of gezondheidsuitkomsten. Het einde van de Tweede Wereldoorlog had een grote impact op de levensomstandigheden en het is daarom verrassend dat dit hoofdstuk geen bewijs vindt voor ouderlijke selectie.

Hoofdstuk drie bestudeert of biologie, en specifiek prenataal testosteron, genderverschillen in educatie kan verklaren. We gebruiken hiervoor exogene variatie in prenataal testosteron dat voortkomt uit een natuurlijk experiment in tweelingen. Prenataal testosteron verplaatst zich in de baarmoeder tussen de mannelijke helft van een tweeling en zijn tweelingbroer of -zus. We houden rekening met de potentiële socialisatie effecten van het opgroeien met een broer of zus door een controlegroep te gebruiken van broertjes en zusjes die maximaal twaalf maanden van elkaar geboren zijn. We vinden dat meisjes met een tweelingbroer, na het controleren voor socialisatie effecten, ongeveer 7% van een standaarddeviatie lager scoren op wiskunde; we vinden geen effect op taal. Dit effect is geconcentreerd onder kinderen die opgroeien in traditionele gezinnen en gebieden. We vermoeden dat conformeren aan de sociale norm een belangrijke rol speelt. Dit betekent dat onze resultaten niet alleen gedreven worden door biologie, maar zich uitten afhankelijk van omgevingsfactoren.

Hoofdstuk 4 bestudeert hoe een aversie tegen de laatste plek schooluitkomsten vormt, en specifiek uitkomsten die belangrijk zijn voor het niveau van het schooltraject op de middelbare school. Dit hoofdstuk analyseert een Nederlandse hervorming die de laagste twee niveaus heeft samengevoegd. Dat betekent dat kinderen die eerst in aanmerking kwamen voor het op één-na-laagste niveau, nu in aanmerking komen voor het laagste niveau. Voor de identificatie gebruik ik dat de hervorming alleen een effect heeft op de kinderen van wie verwacht wordt dat ze naar het laagste niveau gaan. De uitkomstvariabelen zijn het schooladvies van de basisschoolleraar, en de test-score op een belangrijke gestandaardiseerde toets. Mijn bevindingen laten zien dat kinderen van wie verwacht wordt dat ze naar het laagste

niveau gaan een lagere kans hebben om een schooladvies te krijgen boven het laagste niveau, ook scoren ze lager op de gestandaardiseerde toets. De effecten zijn het sterkst voor het zwakste vak van een kind (wiskunde of taal), en in families waar de aversie tegen het laagste niveau vermoedelijk groter is. Een aversie tegen de laatste plek kan zorgen voor slechtere prestaties door een hogere druk om te presteren.

Dit proefschrift bestudeert hoe drie verschillende omstandigheden in de kinderjaren de vorming van menselijk kapitaal beïnvloeden. Het laat zien dat betere sociaaleconomische omstandigheden die zijn voortgekomen uit het einde van de Tweede Wereldoorlog in Nederland niet leiden tot ouderlijke selectie, dat een hogere blootstelling aan prenataal testosteron leidt tot lagere wiskunde scores voor meisjes, en dat aversie tegen de laatste plek kan zorgen voor lagere onderwijsuitkomsten. Het spectrum van omstandigheden in de kinderjaren dat invloed kan hebben op de vorming van menselijk kapitaal is groot, en de drie factoren die onderzocht worden in dit proefschrift vertegenwoordigen slechts een klein deel van de puzzel van hoe iemands familie invloed heeft op iemands levenslot. Echter, een puzzel kan alleen worden opgelost met alle puzzelstukjes in de doos. Dit proefschrift brengt ons een stap dichter bij het begrijpen van welke omstandigheden in de jonge jaren belangrijk zijn voor de vorming van het menselijk kapitaal.

Bibliography

Abdalla, H. I., Burton, G., Kirkland, A., Johnson, M. R., Leonard, T., Brooks, A. A., and Studd, J. W. (1993). Pregnancy: Age, pregnancy and miscarriage: uterine versus ovarian factors. *Human Reproduction*, 8(9):1512–1517.

Abramitzky, R., Delavande, A., and Vasconcelos, L. (2011). Marrying up: the role of sex ratio in assortative matching. *American Economic Journal: Applied Economics*, 3(3):124–57.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics*, 4:1043–1171.

Almond, D. (2006). Is the 1918 influenza pandemic over? long-term effects of in utero influenza exposure in the post-1940 us population. *Journal of Political Economy*, 114(4):672–712.

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

Almond, D. and Currie, J. (2011b). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives*, 25(3):153–72.

Almond, D., Currie, J., and Duque, V. (2018). Childhood circumstances and adult outcomes: Act II. *Journal of Economic Literature*, 56(4):1360–1446.

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. *Annual Review of Economics*, 4(1):185–223.

Ananat, E. O., Gruber, J., Levine, P. B., and Staiger, D. (2009). Abortion and selection. *The Review of Economics and Statistics*, 91(1):124–136.

Ananat, E. O. and Hungerman, D. M. (2012). The power of the pill for the next generation: oral contraception's effects on fertility, abortion, and maternal and child characteristics. *Review of Economics and Statistics*, 94(1):37–51.

Ananth, C. V., Wilcox, A. J., Savitz, D. A., Bowes, W. A., and Luther, E. R. (1996). Effect of maternal age and parity on the risk of uteroplacental bleeding disorders in pregnancy. *Obstetrics & Gynecology*, 88(4):511–516.

Angrist, J. (2002). How do sex ratios affect marriage and labor markets? evidence from America's second generation. *The Quarterly Journal of Economics*, 117(3):997–1038.

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.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

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

Arts, K. and Otten, F. (2013). Stijgende arbeidsparticipatie en minder uittreding bij ouderen. *Sociaaleconomische Trends*, (2013/4).

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. *NBER Working Paper No. 22267*.

Autor, D. H. (2003). Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of Labor Economics*, 21(1):1–42.

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.

Bailey, M. J. (2006). More power to the pill: the impact of contraceptive freedom on women's life cycle labor supply. *The Quarterly Journal of Economics*, 121(1):289–320.

Bailey, M. J. (2013). Fifty years of family planning: new evidence on the long-run effects of increasing access to contraception. *NBER Working Paper No. w19493*.

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.

Becker, G. S. (1960). An economic analysis of fertility. In *Demographic and Economic Change in Developed Countries*, pages 209–240. Columbia University Press.

Becker, G. S., Lewis, H. G., et al. (1973). On the interaction between the quantity and quality of children. *Journal of Political Economy*, 81(2):S279–88.

Becker, G. S. and Tomes, N. (1986). Human capital and the rise and fall of families. *Journal of Labor Economics*, 4(3, Part 2):S1–S39.

Beets, G. (2011). De geboortepiek van 1946: het begin van Nederlands grootste baby-boom wordt 65 jaar. *Demos: Bulletin over Bevolking en Samenleving*, 27.

Ben-Porath, Y. (1973). Economic analysis of fertility in Israel: point and counterpoint. *Journal of Political Economy*, 81(2, Part 2):S202–S233.

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.

Bethmann, D. and Kvasnicka, M. (2013). World war II, missing men and out of wedlock childbearing. *The Economic Journal*, 123(567):162–194.

Bethmann, D. and Kvasnicka, M. (2014). War, marriage markets, and the sex ratio at birth. *The Scandinavian Journal of Economics*, 116(3):859–877.

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

Betts, J. R. (2011). The economics of tracking in education. *Handbook of the Economics of Education*, 3:341–381.

Bhalotra, S. R., Clarke, D., et al. (2016). The twin instrument. *IZA Discussion Paper No. 10405*.

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.

Björklund, A. and Salvanes, K. G. (2011). Education and family background: Mechanisms and policies. *Handbook of the Economics of Education*, 3:201–247.

Black, S. E., Devereux, P. J., et al. (2011). Recent developments in intergenerational mobility. *Handbook of Labor Economics*, 4:1487–1541.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2005). The more the merrier? the effect of family size and birth order on children's education. *The Quarterly Journal of Economics*, pages 669–700.

Blanchflower, D. G. and Oswald, A. J. (2004). Well-being over time in Britain and the USA. *Journal of Public Economics*, 88(7-8):1359–1386.

Blau, F. D. and Kahn, L. M. (2000). Gender differences in pay. *NBER Working Paper No. 7732*.

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? *NBER Working Paper No. 23816*.

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

Bollen, H. and Vroemen, P. (1994). *Canadezen in actie. Nederland najaar '44-voorjaar'46*. Warnsveld.

Borghans, L., Golsteyn, B. H., and Zölitz, U. (2015). Parental preferences for primary school characteristics. *The BE Journal of Economic Analysis & Policy*, 15(1):85–117.

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.

Bound, J. and Turner, S. (2007). Cohort crowding: How resources affect collegiate attainment. *Journal of Public Economics*, 91(5-6):877–899.

Bratti, M. and Cavalli, L. (2014). Delayed first birth and new mothers' labor market outcomes: Evidence from biological fertility shocks. *European Journal of Population*, 30(1):35–63.

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

Brenøe, A. A. and Lundberg, S. (2018). Gender gaps in the effects of childhood family environment: Do they persist into adulthood? *European Economic Review*, 109:42–62.

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.

Brunello, G. (2010). The effects of cohort size on European earnings. *Journal of Population Economics*, 23(1):273–290.

Burgess, S., Greaves, E., Vignoles, A., and Wilson, D. (2015). What parents want: School preferences and school choice. *The Economic Journal*, 125(587):1262–1289.

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.

Card, D., Mas, A., Moretti, E., and Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *American Economic Review*, 102(6):2981–3003.

CBS (1975). Buitenechtelijke geboorten 1840–1973. *Staatsuitgeverij,'s-Gravenhage*.

CBS (2012). Babyboomers, indrukken uit de statistiek. *Den Haag/Heerlen: Centraal Bureau voor de Statistiek*.

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.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, 129(4):1553–1623.

Chevalier, A. and Marie, O. (2017). Economic uncertainty, parental selection, and children's educational outcomes. *Journal of Political Economy*, 125(2):393–430.

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. *Proceedings of the National Academy of Sciences*, 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.

Cunha, F. and Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97(2):31–47.

Currie, J. (2011). Inequality at birth: Some causes and consequences. *American Economic Review*, 101(3):1–22.

Currie, J., Duque, V., and Garfinkel, I. (2015). The great recession and mothers' health. *The Economic Journal*, 125(588).

Dahl, G. B. and Gielen, A. C. (2018). Intergenerational spillovers in disability insurance. *NBER Working Paper No. 24296*.

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

De Graaf, A. and Keil, I. (2001). Werkende moeders [working mothers]. *Index*, 1:10–11.

Dehejia, R. and Lleras-Muney, A. (2004). Booms, busts, and babies' health. *The Quarterly Journal of Economics*, 119(3):1091–1130.

Del Bono, E., Weber, A., and Winter-Ebmer, R. (2012). Clash of career and family: Fertility decisions after job displacement. *Journal of the European Economic Association*, 10(4):659–683.

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.

Dodde, N. L. (1983). *Het Nederlandse onderwijs verandert: ontwikkelingen sinds 1800*. Coutinho.

Doepke, M., Hazan, M., and Maoz, Y. D. (2015). The baby boom and World War II: A macroeconomic analysis. *The Review of Economic Studies*, 82(3):1031–1073.

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.

Ekamper, P., Bijwaard, G., van Poppel, F., and Lumey, L. (2017). War-related excess mortality in The Netherlands, 1944–45: new estimates of famine-and non-famine-related deaths from national death records. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50(2):113–128.

Elias, S. G., van Noord, P. A., Peeters, P. H., den Tonkelaar, I., Kaaks, R., and Grobbee, D. E. (2007). Menstruation during and after caloric restriction: the 1944–1945 Dutch famine. *Fertility and Sterility*, 88(4):1101–1107.

Ellison, G. and Swanson, A. (2009). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American Mathematics Competitions. *NBER Working Paper No. 15238*.

Engelen, T. L. (2005). Kerk en kindertal. over de kracht van religie in de Nederlandse samenleving, 1900-1950. *Nijmegen: Valkhof Pers*.

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.

Fergusson, D. M. and Woodward, L. J. (1999). Maternal age and educational and psychosocial outcomes in early adulthood. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 40(3):479–489.

Feron, E., Schils, T., and Ter Weel, B. (2016). Does the teacher beat the test? the value of the teacher's assessment in predicting student ability. *De Economist*, 164(4):391–418.

Figlio, D. N. and Lucas, M. E. (2004). Do high grading standards affect student performance? *Journal of Public Economics*, 88(9-10):1815–1834.

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. *NBER Working Paper No. 16546*.

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.

Gelber, A. and Isen, A. (2013). Children's schooling and parents' behavior: Evidence from the head start impact study. *Journal of Public Economics*, 101:25–38.

Gianaroli, L., Magli, M. C., Ferraretti, A. P., and Munne, S. (1999). Preimplantation diagnosis for aneuploidies in patients undergoing in vitro fertilization with a poor prognosis: identification of the categories for which it should be proposed. *Fertility and Sterility*, 72(5):837–844.

Gielen, A. and Zwiers, E. (2018). Biology and the gender gap in educational performance: The role of prenatal testosterone in test scores. *IZA Discussion Paper No. 11936*.

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. and Katz, L. F. (2002). The power of the pill: Oral contraceptives and women's career and marriage decisions. *Journal of Political Economy*, 110(4):730–770.

Goldin, C. and Katz, L. F. (2007). The race between education and technology: the evolution of us educational wage differentials, 1890 to 2005. *National Bureau of Economic Research Working Paper No. 12984*.

Goldin, C., Katz, L. F., and Kuziemko, I. (2006). The homecoming of American college women: The reversal of the college gender gap. *Journal of Economic Perspectives*, 20(4):133–156.

Gronau, R. (1977). Leisure, home production, and work–the theory of the allocation of time revisited. *Journal of Political Economy*, 85(6):1099–1123.

Gruber, J., Levine, P., and Staiger, D. (1999). Abortion legalization and child living circumstances: who is the "marginal child"? *The Quarterly Journal of Economics*, 114(1):263–291.

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.

Hastings, J. S., Kane, T. J., and Staiger, D. O. (2005). Parental preferences and school competition: Evidence from a public school choice program. *NBER Working Paper No. 11805*.

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

Heckman, J. J. and Mosso, S. (2014). The economics of human development and social mobility. *Annual Review of Economics*, 6(1):689–733.

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.

Hofstee, S. (2012). *Van periodieke onthouding naar pil. De rol van de R.-K. Artsenvereniging en haar leden in het anticonceptiedebat in de periode 1945-1970 in Nederland*. Doctoral Thesis. Radboud Universiteit Nijmegen.

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.

Hoover-Dempsey, K. V. and Sandler, H. M. (1997). Why do parents become involved in their children's education? *Review of Educational Research*, 67(1):3–42.

Houtenville, A. J. and Conway, K. S. (2008). Parental effort, school resources, and student achievement. *Journal of Human Resources*, 43(2):437–453.

Huber, C. A., Szucs, T. D., Rapold, R., and Reich, O. (2013). Identifying patients with chronic conditions using pharmacy data in Switzerland: an updated mapping approach to the classification of medications. *BMC public health*, 13(1):1030.

Inspectie van het Onderwijs (2014). De kwaliteit van het basisschooladvies: een onderzoek naar de totstandkoming van het basisschooladvies en de invloed van het basisschooladvies op de verdere schoolloopbaan. *Utrecht: Inspectie van het Onderwijs*.

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.

Kamphuis, F., Mulder, L., Vierke, H., Overmaat, M., and Koopman, P. (n.d.). De relatie tussen prima-toetsen en toetsen uit het cito-leerlingvolgsysteem. *Cito, ITS, SCO*.

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.

Kleinhout, G. W. H. (2006). *Jazz als probleem. Receptie en acceptatie van de jazz in de wederopbouwperiode van Nederland 1945-1952*. Doctoral Thesis. Utrecht University.

Klemann, H. A. (2002). *Nederland 1938-1948: economie en samenleving in jaren van oorlog en bezetting*. Boom Koninklijke Uitgevers.

Kloosterman, R. and de Graaf, P. M. (2010). Non-promotion or enrolment in a lower track? the influence of social background on choices in secondary education for three cohorts of Dutch pupils. *Oxford Review of Education*, 36(3):363–384.

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

Krimp, R. and Kemperman, J. (2015). *De doden tellen: slachtofferaantallen van de Tweede Wereldoorlog en sindsdien*. Nationaal Comité 4 en 5 mei.

Kuziemko, I., Buell, R. W., Reich, T., and Norton, M. I. (2014). "Last-place aversion": Evidence and redistributive implications. *The Quarterly Journal of Economics*, 129(1):105–149.

Leuven, E., Lindahl, M., Oosterbeek, H., and Webbink, D. (2010). Expanding schooling opportunities for 4-year-olds. *Economics of Education Review*, 29(3):319–328.

Lindeboom, M. and Van Ewijk, R. (2015). Babies of the war: Effect of war exposure early in life on mortality throughout life. *Biodemography and Social Biology*, 61(2):167–186.

Lindo, J. M. (2010). Are children really inferior goods? evidence from displacement-driven income shocks. *Journal of Human Resources*, 45(2):301–327.

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.

Lumey, L. H., Stein, A. D., and Susser, E. (2011). Prenatal famine and adult health. *Annual Review of Public Health*, 32:237–262.

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.

Luttmer, E. F. (2005). Neighbors as negatives: Relative earnings and well-being. *The Quarterly Journal of Economics*, 120(3):963–1002.

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*, 322(5906):1331–1332.

Mamelund, S.-E. (2006). A socially neutral disease? individual social class, household wealth and mortality from Spanish influenza in two socially contrasting parishes in kristiania 1918–19. *Social Science & Medicine*, 62(4):923–940.

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, A. R. (2009). Motherhood delay and the human capital of the next generation. *American Economic Review*, 99(2):154–58.

Miller, A. R. (2011). The effects of motherhood timing on career path. *Journal of Population Economics*, 24(3):1071–1100.

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.

Ministry of Education (2005). Het vmbo: beelden, feiten en toekomst. interdepartementaal beleidsonderzoek 2004-2005. *Ministerie van Onderwijs, Cultuur en Wetenschap*.

Mølland, E. (2016). Benefits from delay? the effect of abortion availability on young women and their children. *Labour Economics*, 43:6–28.

Myers, C. K. (2017). The power of abortion policy: Reexamining the effects of young women's access to reproductive control. *Journal of Political Economy*, 125(6):2178–2224.

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.

Nobles, J., Frankenberg, E., and Thomas, D. (2015). The effects of mortality on fertility: Population dynamics after a natural disaster. *Demography*, 52:15–38.

Nollenberger, N., Rodriguez Planas, N., and Sevilla Sanz, A. (2014). The math gender gap: The role of culture. *IZA Discussion Paper No. 8379*.

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.

Okkema, B. (2012). *Trees krijgt een Canadees: bevrijdingskinderen in Nederland*. Walburg Pers b.v., Zutphen.

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.

Pellicer, A., Simón, C., and Remohí, J. (1995). Effects of aging on the female reproductive system. *Human Reproduction*, 10(suppl_2):77–83.

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.

Perry, C. (2004). How do female earnings affect fertility decisions? *Massachusetts Institute of Technology, PhD Thesis*.

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.

Porter, M. (2016). How do sex ratios in china influence marriage decisions and intra-household resource allocation? *Review of Economics of the Household*, 14(2):337–371.

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.

Rau, T., Sarzosa, M., and Urzúa, S. S. (2017). The children of the missed pill. *NBER Working Paper 23911*.

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.

Roeleveld, J. and Vierke, H. (2003). Uitval en instroom bij de derde meting van het prima-cohortonderzoek. *Amsterdam/Nijmegen: SCO-Kohnstamm Instituut/ITS*.

Rønning, M. (2011). Who benefits from homework assignments? *Economics of Education Review*, 30(1):55–64.

Roseboom (2010). *Baby's van de Hongerwinter. De onvermoede erfenis van ondervoeding*. Augustus.

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.

Royer, H. (2004). What all women (and some men) want to know: Does maternal age affect infant health? *Center for Labor Economics, University of California, WORKING PAPER NO. 68*.

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

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

Schaller, J. (2016). Booms, busts, and fertility. *Journal of Human Resources*, 51(1).

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.

Schuyt, K. and Taverne, E. (2004). *Dutch Culture in a European Perspective Vol 4; 1950: Prosperity and welfare*. Palgrave Macmillan, UK.

Sijes, B. A. (1990). *De arbeidsinzet: de gedwongen arbeid van Nederlanders in Duitsland, 1940-1945*. Number Nr. 11. Netherlands State Insitute for War Documentation, Amsterdam.

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.

Stein, Z. and Susser, M. (2000). The risks of having children in later life. *Western Journal of Medicine*, 173(5):295.

Stein, Z., Susser, M., Saenger, G., and Marolla, F. (1975). Famine and human development: The Dutch hunger winter of 1944-1945.

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).

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.

Timmermans, A., Kuyper, H., and Van der Werf, G. (2013). Schooladviezen en onderwijsloopbanen. voorkomen, risicofactoren, en gevolgen van onder- en overadvisering. *Groningen: GION*.

Torun, H. and Tumen, S. (2016). The empirical content of season-of-birth effects: An investigation with Turkish data. *IZA Discussion Paper No. 10203*.

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 Bavel, J. and Reher, D. S. (2013). The baby boom and its causes: What we know and what we need to know. *Population and Development Review*, 39(2):257–288.

Van Boxtel, H., Engelen, R., and De Wijs, A. (2011). Wetenschappelijke verantwoording van de eindtoets basisonderwijs 2010. *Arnhem: Cito*.

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.

Van den Brink, T. (1950). Birth rate trends and changes in marital fertility in the Netherlands after 1937. *Population Studies*, 4(3):314–332.

Van der Brakel, M. and Moonen, L. (2013). Inkomensmobiliteit tussen generaties relatief hoog in Nederland. *Sociaaleconomische trends, April, Centraal Bureau voor de Statistiek*.

Van Ewijk, R. and Lindeboom, M. (2017). Why people born during World War II are healthier. *Gutenberg School of Management and Economics Discussion paper 1619.*

Van Poppel, F. and Willekens, F. (1982). The decrease in the age at first marriage in the Netherlands after the Second World War: a log-linear analysis. *Voorburg Netherlands Netherlands Interuniversity Demographic Institute 1982 Jun.*

Van Poppel, F. W. (1985). Late fertility decline in the Netherlands: The influence of religious denomination, socioeconomic group and region. *European Journal of Population/Revue Européenne de Démographie*, 1(4):347–373.

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). *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*, 21(8):1069–1071.

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).

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

690 S. SINGH, *Three Essays on the Insurance of Income Risk and Monetary Policy*

691 E. SILDE, *The Econometrics of Financial Comovement*

692 G. DE OLIVEIRA, *Coercion and Integration*

693 S. CHAN, *Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation*

694 P. GAL, *Essays on the role of frictions for firms, sectors and the macroeconomy*

695 Z. FAN, *Essays on International Portfolio Choice and Asset Pricing under Financial Contagion*

696 H. ZHANG, *Dealing with Health and Health Care System Challenges in China: Assessing Health Determinants and Health Care Reforms*

697 M. VAN LENT, *Essays on Intrinsic Motivation of Students and Workers*

698 R.W. POLDERMANS, *Accuracy of Method of Moments Based Inference*

699 J.E. LUSTENHOUWER, *Monetary and Fiscal Policy under Bounded Rationality and Heterogeneous Expectations*

700 W. HUANG, *Trading and Clearing in Modern Times*

701 N. DE GROOT, *Evaluating Labor Market Policy in the Netherlands*

702 R.E.F. VAN MAURIK, *The Economics of Pension Reforms*

703 I. AYDOGAN, *Decisions from Experience and from Description: Beliefs and Probability Weighting*

704 T.B. CHILD, *Political Economy of Development, Conflict, and Business Networks*

705 O. HERLEM, *Three Stories on Influence*

706 J.D. ZHENG, *Social Identity and Social Preferences: An Empirical Exploration*

707 B.A. LOERAKKER, *On the Role of Bonding, Emotional Leadership, and Partner Choice in Games of Cooperation and Conflict*

708 L. ZIEGLER, *Social Networks, Marital Sorting and Job Matching. Three Essays in Labor Economics*

709 M.O. HOYER, *Social Preferences and Emotions in Repeated Interactions*

710 N. GHEBRIHIWET, *Multinational Firms, Technology Transfer, and FDI Policy*

711 H.FANG, *Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing*

712 Y. KANTOR, *Urban Form and the Labor Market*

713 R.M. TEULINGS, *Untangling Gravity*

714 K.J.VAN WILGENBURG, *Beliefs, Preferences and Health Insurance Behavior*

715 L. SWART, *Less Now or More Later? Essays on the Measurement of Time Preferences in Economic Experiments*

716 D. NIBBERING, *The Gains from Dimensionality*

717 V. HOORNWEG, *A Tradeoff in Econometrics*

718 S. KUCINSKAS, *Essays in Financial Economics*

719 O. FURTUNA, *Fiscal Austerity and Risk Sharing in Advanced Economies*

720 E. JAKUCIONYTE, *The Macroeconomic Consequences of Carry Trade Gone Wrong and Borrower Protection*

721 M. LI, *Essays on Time Series Models with Unobserved Components and Their Applications*

722 N. CIURILÃ, *Risk Sharing Properties and Labor Supply Disincentives of Pay-As-You-Go Pension Systems*

723 N.M. BOSCH, *Empirical Studies on Tax Incentives and Labour Market Behaviour*

724 S.D. JAGAU, *Listen to the Sirens: Understanding Psychological Mechanisms with Theory and Experimental Tests*

725 S. ALBRECHT, *Empirical Studies in Labour and Migration Economics*

726 Y.ZHU, *On the Effects of CEO Compensation*

727 S. XIA, *Essays on Markets for CEOs and Financial Analysts*

728 I. SAKALAUSKAITE, *Essays on Malpractice in Finance*

729 M.M. GARDBERG, *Financial Integration and Global Imbalances*

730 U. THÜMMEL, *Of Machines and Men: Optimal Redistributive Policies under Technological Change*

731 B.J.L. KEIJSERS, *Essays in Applied Time Series Analysis*

732 G. CIMINELLI, *Essays on Macroeconomic Policies after the Crisis*

733 Z.M. LI, *Econometric Analysis of High-frequency Market Microstructure*

734 C.M. OOSTERVEEN, *Education Design Matters*

735 S.C. BARENDSE, *In and Outside the Tails: Making and Evaluating Forecasts*

736 S. SÓVÁGÓ, *Where to Go Next? Essays on the Economics of School Choice*

737 M. HENNEQUIN, *Expectations and Bubbles in Asset Market Experiments*

738 M.W. ADLER, *The Economics of Roads: Congestion, Public Transit and Accident Management*

739 R.J. DÖTTLING, *Essays in Financial Economics*