Essays in empirical microeconomics

Péter, A.N.

Publication date
2016

Document Version
Final published version

Link to publication

Citation for published version (APA):
Péter, A. N. (2016). *Essays in empirical microeconomics*. [Thesis, fully internal, Universiteit van Amsterdam].

tinbergen
institute

Noémi Péter

Essays in Empirical Microeconomics

Universiteit van Amsterdam

ESSAYS IN EMPIRICAL MICROECONOMICS

ISBN 978 90 3610 458 6

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

This book is no. 658 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.

ESSAYS IN EMPIRICAL MICROECONOMICS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
Prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op donderdag 22 september 2016, te 12:00 uur

door

Anna Noémi Péter

geboren te Cluj-Napoca/Kolozsvár, Roemenië

PROMOTIECOMMISSIE

Promotor:

Prof. dr. H. Oosterbeek, Universiteit van Amsterdam

Overige leden:

Dr. A.S. Booij, Universiteit van Amsterdam
Prof. dr. ir. J.C. van Ours, Tilburg University
Prof. dr. E.J.S. Plug, Universiteit van Amsterdam
Prof. dr. A.R. Soetevent, Rijksuniversiteit Groningen
Prof. dr. J.H. Sonnemans, Universiteit van Amsterdam

Faculteit Economie en Bedrijfskunde

Acknowledgments

As my defense is approaching, I would like to take the opportunity to express my gratitude towards the people who helped me get to this point. Let me start with my supervisor. Hessel, thank you for believing in me and supporting me all along. You stood behind me from the first moment, when I applied to the UvA to do my PhD. You also trusted me in that you let me pursue my own research topics. While you gave me a lot of freedom, you were always accessible if I wanted to ask you for guidance. You shaped my thinking about virtually every element of the profession, let it be the way I formulate a research question, analyze data, write a paper or do a presentation. You also gave valuable advice on some non-professional topics, which I appreciate just as much.

The other two persons who were somewhat of a mentor-figure during my PhD were Monique and Erik. Monique, thank you for all the advice and help. I learned a lot from you in general, and of course about the bounds method in particular. You definitely left a mark, especially on Chapter 4 of this thesis. Erik, thanks for your insights, guidance and support. When I encountered difficulties, you often helped me find a good path so that I could proceed forward.

I would also like to express my gratitude towards my co-authors. Thomas, I enjoyed our joint work on the multitasking paper and it was also great to go through the publication process together. You inspire me in many ways. Among other things, I am amazed by your ability to express complex ideas so clearly and concisely. I learned a great deal from you during our joint work, and I am glad that we continue our collaboration beyond the multitasking paper. Dinand and Petter, I am grateful for the collaboration on the twin sibling gender study. Working on this project was a valuable experience and I learned a lot from our fruitful discussions. It was very beneficial to get insights from such experienced colleagues. At this point I would like to thank Stefan as well, with whom I work on a project that is not part of this thesis. Stefan, I am very glad of the opportunity to work with you. I find the project very exciting and I am enthusiastic about the possibilities that it still holds for us.

I was lucky to have had many other great people around me during my PhD. Adam, Nienke, Nadine, José, Ferry, Jona, Stephen, Diana, Sabine, Lenny, Lydia, thank you

for your help, advice and company. I enjoyed having such a friendly and enriching atmosphere in the human capital group. I am also grateful to the colleagues from the VU with whom we joined forces for the work-in-progress seminars: Bas, Jonneke, Sandor, Sandra, Nynke, Mathilde, Sabien, Elisabeth. These seminars provided so much insights and valuable feedback, I am very happy that I could be part of this. I would also like to thank the CREED people in general, and Roel, Arthur, Joep and Thomas de Haan in particular, for being so helpful and supportive of my experimental pursuits. There were many other colleagues as well who contributed to a pleasant environment at the UvA; I want to thank especially Sylvia, Sander, András, Martin and Andrej at this point.

Special thanks go to my paranymphs. Nienke and Anita, I am grateful for your close friendship. I really like that we can share so many things with each other. Your advice and support helped me tremendously during pregnancy and after childbirth, and it also helped me get through some other difficult personal situations. I am glad that we trust each other so much. I enjoy the time spent together and it is great that we have joint family programs as well. Thank you for all this, I hope that there will be many more such memorable moments to come.

I am also very grateful to the bigger "Hungarian lunch" group. Dávid, Anita, Pisti, Böbi, Sanyi, Andris and Orsi, thank you for the fun times and for all the help with respect to moving and children. I am happy that we have such a nice social group and it is really good to see how cutely our kids play together.

I would also like to thank my friends who are in Hungary, especially Eszter, Zsuzsi and Gyöngyi. We might live far away from each other, but there are many things that connect us. I am glad that we did not lose touch during all these years and I really enjoy our reunions. It helps me recharge a bit when we revisit the many great memories, and it is good to see that we can still immerse in conversations so deeply that we hardly recognize how much time has passed by.

Most of all, I would like to thank my family. I have to start this with my family of origin. Mom and Dad, I know how hard both of you worked, often double shifts, to provide good opportunities for your children. Whenever I asked for an investment in my education, you were always there for me. I cannot thank you enough for that. I would not have been able to reach this much in life without your continuous support. Thanks go to my siblings as well. András and Imola, thank you for helping out, especially when we were preparing for the birth of Dani. Imola, thank you also for all the advice and support that you gave with respect to childbirth and child care. I also want to thank your family at this point: Botond, Ábel, Eszter and Nóri, thank you for your kindness and for all the wonderful moments that we shared. Your interactions with Dani often melt my heart, and I am very happy that there is such a special bond between us.

I am also grateful to my extended family, particularly Matyi and Gabi. You were big fans of my PhD, showed a real interest in my research and always encouraged me. Our relationship has always been warm and close, not only because I am Péter's wife but also because of the emotional connection between us. I hope that it has always been clear to you how much this means to me. Gabi, thank you for your continuous support. Matyi, I wish I could share these moments with you, I know that you would be very happy and proud if you were still among us.

Last, but certainly not least, I would like to thank my amazing husband and son, Péter and Dani. Thank you for your love, support, patience and encouragement. Thank you for all the joyful moments. Péter, thank you for being such a wonderful partner, in every possible sense of the word. We have been through good times and tough times, and we were always there for each other. I know that I can always count on your support, and you can always count on mine. This not only helped me to get through difficulties during my PhD, but it also gives me a lot of confidence in general. I know that whatever challenge life might throw at us, we will meet it together. I love you, and I cannot wait to see what the future still holds for us.

Contents

Chapter 1

Introduction

In empirical microeconomics, the identification of causal effects has a central role. It is a challenging undertaking because correlations between two variables can arise not only because of a causal relationship but also due to other reasons. A typical problem is non-random selection. This can arise due to self-selection: individuals with different abilities or preferences might choose different treatments[1] for themselves. It can also be that the selection is made by someone else, for example, if the government prescribes a specific treatment for a part of the population based on some background characteristics. Both type of non-random selection implies that differences between the outcomes of individuals in different treatments can arise not only due to the difference in treatments but also due to differences in other factors, such as abilities, preferences or other background characteristics.

One of the main tasks of empirical economists is to address such selection issues and find ways in which we can disentangle the causal effect of a treatment from other confounding factors. In some contexts, it is possible to run an experiment, where the researcher can allocate individuals to treatments randomly. This random assignment means that on average, we can expect individuals' background characteristics to be similar across treatments. As a result, we can interpret the difference between the outcomes as the causal effect of the difference in treatment.

While such controlled experiments provide a very clean setup to isolate causal effects, they are not always feasible. Many interesting research questions are centered around treatments that cannot be allocated randomly, for example because of budgetary, ethical or political considerations. Therefore researchers are constantly trying to find other ways to measure treatment effects. As we explained above, random assignment is the key to identification in controlled experiments. Thus, if we can find

[1] We follow the economics jargon and use the word "treatment" to refer to the explanatory variable of interest. A "treatment" does not necessarily mean a medical treatment; for example, a researcher who is interested in the effect of an educational program may refer to this educational program as the "treatment".

this component occurring naturally in a non-controlled setting, we can exploit it to identify the treatment effect. Such “natural experiments” are therefore very valuable for empirical research.

But what if there is no experiment, neither controlled nor natural, that could be used to evaluate the effect of a specific treatment? Do we have to give up the hope on causal research in those cases? Not necessarily. When strong conditions such as independence are not satisfied, one could still examine what can be said about the causal effect under weaker conditions. For example, Manski (1989) has shown that if the outcome variable is bounded, it is possible to construct upper and lower bounds around the treatment effect. These bounds could be tightened if we can successively layer various nonparametric assumptions on them. Thus, while we cannot point identify the effect, we might still be able to establish informative bounds around the effect.

As we can see, there are various methods that can be used for research on causal effects. Without giving a complete overview, we discussed three techniques: controlled experiments, natural experiments and nonparametric bounds. These are the three methods that the three studies in this thesis apply. The reason for this methodological diversity is that I viewed my PhD study as an opportunity to build a broad research toolkit. My motivation behind this objective was twofold. First, I think that looking at identification problems from different angles helps to deepen knowledge at a conceptual level. Second, I think that a broad toolkit can give more freedom in the choice of topics, because then there is no constraint to work only on questions that can be answered with a particular method.

Each of the abovementioned methods are applied in one chapter. Chapter 2 uses a lab experiment to examine the productivity effects of different work schedules. Chapter 3 exploits a natural experiment to examine how the gender of a sibling affects individuals’ education, earnings and family formation. Chapter 4 uses the abovementioned nonparametric bound technique to examine how additional school resources affect students’ test scores in a Weighted Student Funding system. While the three chapters use different methods, a common feature is that they all try to tackle some kind of a selection problem. In addition, a common theme of these studies is that they look at individual outcomes that can be of interest to labor economists in a broad sense, such as productivity, earnings, family formation and educational outcomes. In the remainder of this chapter I will briefly introduce the three studies in turn.

Chapter 2 focuses on multitasking, which we define as a work schedule in which the worker is switching back and forth between two ongoing tasks. We examine four research questions related to multitasking. First, we examine how multitasking affects productivity. Second, we examine whether individuals optimally choose their degree of multitasking or whether they perform better under an externally imposed schedule.

Third, we examine whether there are gender differences in the effect of multitasking on productivity. And fourth, we examine whether there are gender differences in the propensity to multitask. The first pair of questions is motivated by practical concerns: should employers impose a schedule on employees, and if so, what kind of a schedule? The second pair of questions is motivated by the gap between popular views and scientific evidence: it is widely held that women are better at multitasking, while scientific evidence on the subject is scarce.

It would be difficult to examine the abovementioned research questions with data from a non-controlled environment, because individuals who follow different work schedules are probably different in many regards, for example, in their abilities, preferences, motivations and external constraints. Therefore we conducted a lab experiment to answer our research questions. In the experiment subjects are randomly allocated to different work schedules. They have to perform two separate tasks according to one of three different treatments: one where they perform the tasks sequentially, one where they are forced to multitask, and one where they can freely choose their schedule. We kept the amount of time per task constant across treatments to ensure that performance differences between treatments measure the productivity effect of the different schedules.

We find that subjects who are forced to multitask perform significantly worse than those forced to work sequentially. Surprisingly, subjects who can freely organize their own schedule also perform significantly worse. These results suggest that scheduling is a significant determinant of productivity. Finally, our results do not support the stereotype that women are better at multitasking. Women suffer as much as men when forced to multitask and are actually less inclined to multitask when being free to choose.

The research question of Chapter 3 is how the gender of a sibling affects individual's earnings, education and family formation. This question was inspired by the observation that close family members such as siblings have a big potential to affect people's lives. The motivation to look specifically at the gender of the sibling is that having a brother implies a different family environment than having a sister. It is well established that men and women (and hence brothers and sisters) are different in many regards and they are also treated differently by parents. Thus, those with a brother are exposed to different family conditions than those with a sister, which may have an impact on their labor market outcomes and family formation. We look at these outcomes in particular because many gender differences relate to these domains.

Although children cannot choose the gender of their siblings, the identification of the treatment effect becomes challenging because of parental preferences. If parents prefer certain sex compositions over others, children's gender affects not only the outcomes of other children but also the existence of potential additional children. Since parental preferences may affect children's outcomes as well, a selection bias can arise, which is similar in nature to the bias in Heckman (1979). We circumvent the selection problem by exploiting a natural experiment: dizygotic twins. In these cases, the two children are born at the same time, so parents cannot make decisions about one twin based on the gender of the other twin.

We find that both men and women are influenced by the gender of their sibling, but in a different way. Men with brothers earn more and are more likely to get married and have children than men with sisters. Women with sisters obtain lower education and give birth earlier than women with brothers. Our analysis shows that the most likely explanation for these findings is that siblings affect each other via various social mechanisms.

Chapter 4 examines how test scores are affected by additional school resources that are provided in the Dutch Weighted Student Funding (WSF) system. WSF is a school finance policy that aims at improving the educational outcomes of disadvantaged students. It is characterized by three main elements: 1) a money-follows-students system whereby the funding of a school depends on the number of students, 2) additional weights are assigned to disadvantaged students and therefore their schools get more resources, and 3) schools are free to use these resources as they want. In the Dutch version of the WSF that we examine, disadvantaged status is defined by parental background.

The identification of the effect of the extra funding is clearly of high policy interest. However, it is difficult to disentangle this causal effect from other factors. The selection problem at hand is caused by the fact that the assignment of the extra funding is based on family background. This non-random assignment implies that differences in the performance of the treated and the non-treated group can arise not only due to the funding but also due to the difference in family background.

To isolate the causal impact of the extra funding, we use the nonparametric partial identification method that was developed in Manski (1989), Manski (1997), Manski and Pepper (2000) and Manski and Pepper (2009). We start without imposing assumptions to obtain worst-case bounds and then we layer various nonparametric assumptions to tighten the bounds. We make use of three types of assumptions: Monotone Treatment Selection, Monotone Instrumental Variable and Monotone Treatment Response. For the MIV bounds, we use average neighborhood income, thereby assuming a non-negative relationship between mean potential test scores and average income in the neighborhood. The MTS-MIV bounds indicate that the extra funding has a positive impact on math scores, which can be potentially very high. The bounds around the impact on language scores, information-processing scores and total scores also allow for large positive impacts, but they also allow for a more moderate negative impact. We subsequently show that adding the stronger MTR assumption tightens the bounds such that all lower bounds are above zero. Thus, we find that when the impact on students is assumed to be non-negative, the average treatment effect seems to be significantly positive.

Chapter 2

Multitasking[1]

2.1 Introduction

Although multitasking[2] is increasingly common in the modern work environment, its productivity effects remain underexplored. Furthermore, the stereotype that women are better at multitasking is almost universally accepted but, again, scientific evidence is missing. This chapter fills these gaps through an experimental design which allows us to answer the following research questions. First, how does multitasking affect productivity? Second, do people perform better when they are allowed to choose their own schedule? Third, are there indeed gender differences in the effect of multitasking on productivity? And fourth, are there gender differences in the propensity to multitask?

The first pair of questions is motivated by a practical concern: how to schedule tasks optimally. Is sequential execution advisable, or is it more productive to alternate (that is, to multitask)? Is it optimal to let workers choose their own schedule or should companies impose one? Although it seems intuitive that scheduling has an impact on productivity, this topic has received little attention so far in economics. The vast literature on multiple tasks focuses instead on the pros and cons of bundling different tasks into a single job and on what kind of tasks should be grouped together.[3] The literature on workers' decision making rights does not address scheduling directly

[1] This chapter is based on Buser and Peter (2012). Financial support from the Speerpunt Behavioural Economics of the University of Amsterdam is gratefully acknowledged.

[2] In this chapter, by multitasking we mean switching back and forth between two ongoing tasks. The concepts of multitasking and task-switching are discussed in more detail in Section 2.2.

[3] Holmstrom and Milgrom (1991) argue that tasks assigned to the same worker should have a similar degree of measurability; otherwise the worker focuses on the task that is easier to measure, leading to a suboptimal allocation of effort across tasks. Schottner (2007) notes that some of the concerns can be mitigated by using relational contracts. Friebel and Yilmaz (2010) show that assigning call centre workers a greater variety of tasks decreases average individual productivity. Drago and Garvey (1998), on the other hand, show that workers with more varied tasks shirk less and help their colleagues more. This result could be due to workers' preference for more versatile work, as suggested by Lindbeck and Snower (2000).

either.[4] The only paper we found analyzing the impact of work schedules is Coviello et al. (2011). They examine court cases, where a natural candidate for the measure of performance is average duration. They find that judges who work on many cases in parallel take more time than judges who work sequentially to complete similar portfolios of cases. Their results confirm that work schedules are indeed an important determinant of productivity.

The second pair of research questions is motivated by the gap between popular views and scientific evidence: best-selling books advertise that women are better at multitasking as a scientifically established fact[5], while in reality this gender difference has not so far been shown by any peer-reviewed paper.[6] While empirical evidence is lacking, the view that women are better at multitasking gets support from the hunter-gatherer hypothesis, a theoretical argument in biological anthropology. In particular, Fisher (1999) claims that the prehistoric division of work "built" different aptitudes into the male and female brain through natural selection. Different skills are required for hunting, performed by males, than for gathering, performed by women. As a consequence, argues Fisher, women think "contextually", as they synthesize many factors into a "web of factors", while men think linearly, focusing on a single task until it is done. This implies that women are both better at multitasking and more inclined to do it. Our design allows us to test both these hypotheses explicitly.

We examine the above research questions empirically by conducting an experiment in which subjects are randomly allocated to different work schedules. Participants have to perform two separate tasks (a Sudoku and a Word Search puzzle) according to one of three different treatments: one where they perform the tasks sequentially, one where they are forced to alternate between the two tasks, and one where they can freely organize their work. The amount of time spent on each task is identical in each treatment. Performance differences between treatments therefore measure the productivity effect of the different schedules. Relative performance in the third treatment, where subjects can freely choose the degree of multitasking, is indicative on whether individuals should be free to organize their own schedule. Gender differences in performance in

[4] This line of research focuses on the trade-off between losing control and utilizing information from the lower levels of hierarchy (see Lazear and Gibbs (2009), Ch. 5). When concrete examples are given, they relate to the selection of projects or ideas that the workers work on (see for example Aghion and Tirole (1997) and Zábojník (2002)).

[5] See for example Pease and Pease (2001) and its adaptation, Why Men Can Only Do One Thing at a Time and Women Never Stop Talking (Pease and Pease, 2003).

[6] We searched extensively for peer-reviewed publications about gender differences in multitasking, but the closest we could find is Criss (2006) and Havel (2004), two undergraduate theses. Both examined subjects who had to perform some specified tasks while tallying keywords from a song/story. None of them found gender differences in productivity when multitasking. Nonetheless, we do not know what these findings mean with respect to multitasking as none of them had a control group. The media regularly mentions research which supposedly shows that women are better at multitasking but to the best of our knowledge, none of this has been published in peer-reviewed journals.

the second treatment allow us to test whether men perform worse than women when they are forced to multitask. Finally, choices in the third treatment are used to test whether men indeed prefer a more sequential schedule than women.

Related to our study is a literature on 'task-switching' in psychology (see Monsell, 2003 for a review). In these experiments, a series of stimuli is presented to participants who have to perform a short task on each stimulus. For example, pairs of numbers are shown and subjects have to either add them up or to multiply them (see Rubinstein et al., 2001). From time to time, the required operation changes. It is commonly found that there are 'switching costs' associated with changing tasks, i.e. the response to the stimuli is slower after a task-switch. This literature can, however, not answer our research questions. The tasks used are too simple to expect any advantages from multitasking and subjects are not allowed to choose their schedule freely. Also, these experiments are not usually incentivized. In contrast, we use two complex tasks of much longer duration. These tasks are contingent, meaning that after switching back, subjects return to working on the same ongoing problem. Subjects can therefore expect an advantage from alternating: they can switch when they get stuck and later look at the same problem with a 'fresh eye'. Indeed, our subjects do switch when they are allowed to.

Finally, none of the psychological experiments are designed to examine gender differences. Their samples are generally too small to do so and often characterized by strong gender imbalances. Our comparatively large and balanced sample, on the other hand, allows us to test both whether there are gender differences in the effects of multitasking and in the propensity to multitask.

2.2 Definitions

There are several possible definitions of multitasking.[7] The variant we address in our experiment is the one that is most relevant in the workplace: people switching between multiple contingent tasks.[8] It is also this form of multitasking which has garnered the most interest in the popular press, where articles about the productivity effects of

[7] Multitasking is often thought of as the performance of multiple tasks at one time, but this definition is at odds with the findings of many psychologists and neuroscientists. Pashler (1994) reviews the related literature and concludes that our ability to simultaneously carry out even simple cognitive operations is very limited. Using brain scanners, Dux et al. (2006) localize a neural network which acts as a central bottleneck of information processing by precluding the selection of response to two different tasks at the same time. Furthermore, Dux et al. (2009) show that while training can increase the speed of information processing in this brain region, it remains true that tasks are not processed simultaneously but in rapid succession. Simultaneity is an illusion, which occurs if the tasks are so simple that the alternations are very quick.

[8] Switching means redirecting attention from one task to another. The reader can find more details about its neurological background in Dux et al. (2006).

multitasking are common. In our experiment, subjects continue working on the same problem after they return from their work on the second task, similar to an employee switching between tasks or having his work at hand interrupted by another, perhaps more urgent task. Another relevant example is when people multitask on a computer, switching back and forth between windows or tabs.

Note that our definition of multitasking is similar to what psychologists call task-switching, but there is an important difference between the two: contingency. When tasks are contingent, there are potential benefits to multitasking, such as seeing an old problem with a 'fresh eye'. In contrast, in previous task-switching experiments subjects get a new stimulus to work on each time (e.g. they get a new pair of numbers to add up), so only the operation remains the same, not the problem they are working on. In this way, we aim to investigate the type of multitasking which occurs in a modern work environment where employees switch between several demanding and ongoing tasks.

In line with this aim, we chose tasks that require primarily mental effort and have virtually no physical aspects. Our tasks are therefore not chosen to resemble household activities like doing the dishes and taking care of children. Research that focuses on such household activities can be found in Kalenkoski and Foster (2010).[9]

Note also that our definition of multitasking explicitly ignores a possible advantage of working on multiple tasks, namely the possibility of reallocating time between these tasks. We see time allocation as separate from multitasking: it is possible to reallocate time between tasks while still executing them sequentially and conversely it is possible to multitask without reallocating time between tasks. One strength of our design is that it clearly separates these two mechanisms: we keep time allocation constant so we can identify the effect of scheduling.

2.3 Experimental design and data

2.3.1 Treatments and groups

Three treatments were applied during the experiment: Treatment Single, Treatment Multi, and Treatment Choice (subjects were randomly allocated to treatments within each session and were unaware of these labels). In Treatment Single, subjects had to work on two tasks consecutively, for 12 minutes each. In Treatment Multi, subjects were forced to switch between the two tasks approximately every four minutes[10], resulting in the same total time constraint per task as before. Subjects did not know how

[9] This research builds on time diaries that allow individuals to list primary and secondary activities in a time interval and investigates correlations between parental time use and child outcomes.

[10] Gonzalez and Mark (2004) found that information workers spend on average 3 minutes on a task without interruption; this average might be somewhat higher in a less fast-paced environment.

many switches would occur and the time intervals between switches varied, making anticipation unlikely. In Treatment Choice, subjects could alternate between the two tasks by pressing a 'Switch' button, subject to the same time constraint per task as before (12 minutes each). A timer informed subjects about the remaining time for each task. When the 12 minutes for one task expired, the screen changed automatically to the other task and the Switch button could not be used anymore.

It is important to see that this design ensures that the same amount of time is spent on each task in all three treatments. If we tried to resemble simultaneity, for example by splitting the screen, we could not determine how much time subjects spend on each task, and therefore we would not know whether performance between treatments differs due to differences in the amount of time allocated to the two tasks or due to differences in the schedules.

As shown in Table 2.1, subjects were assigned to three groups. Every subject played two rounds, the first of which was Treatment Single. In the second round, subjects in Group 1 played Treatment Single again, subjects in Group 2 played Treatment Multi, and subjects in Group 3 played Treatment Choice. The subjects knew from the start that there would be two rounds and that they would work on one Sudoku puzzle and one Word Search puzzle in each. The puzzles given in Round 2 were different from the puzzles in Round 1 (but they were the same for all subjects within rounds).

Table 2.1. Treatments of each group

	Group 1	Group 2	Group 3
Round 1	Single	Single	Single
Round 2	Single	Multi	Choice

This design allows us to answer all four research questions and the fact that Group 1 plays Single twice allows for a difference-in-differences approach. This enables us to correct for learning effects and performance drops due to exhaustion or boredom. To examine the effect of forced multitasking on productivity, we can compare the performance difference between Round 1 and Round 2 of Group 2 to the performance difference of Group 1. To examine the effect of a self-chosen work schedule, we can compare the performance difference of Group 3 to the performance differences of the other two groups. If subjects choose the optimal work schedule, we should see that the performance difference of Group 3 is at least as high as the performance difference of the other two groups.[11]

To examine gender differences in the effects of multitasking on productivity, we follow a difference-in-difference-in-differences approach. Note that any gender differ-

[11] Since subjects in Group 3 can choose whether or not to alternate, finding that they performed worse than the other groups would disprove that they chose optimally.

ence in performance cannot be led by differences in task proficiency since we compare performance in Round 2 to a subject's own performance in Round 1. Besides, Group 1 captures any gender differences in learning or exhaustion. For Group 2, any gender difference in performance therefore can only come from differences in the reaction to multitasking. For Group 3, both the reaction to multitasking and the self-chosen degree of multitasking determine the performance difference.

Finally, to examine whether there is any gender difference in the propensity to multitask, we check whether there is a gender difference in the number of switches in Treatment Choice. The propensity to multitask might vary with proficiency: subjects who perform well might find switching easier or more beneficial. Alternatively, subjects who get stuck more often may want to switch more often. To avoid attributing such effects to gender differences in multitasking, we control for performance in Round 1.

2.3.2 Tasks

Our design requires tasks that are not gender-specific and for which multitasking is natural and possibly beneficial. For these reasons, we have chosen Sudoku and Word Search as tasks.[12] Sudoku is played over a 9x9 grid, divided into 3x3 sub-grids called "regions". The left panel of Figure 2.1 illustrates that a Sudoku puzzle begins with some of the grid cells already filled with numbers. The objective of Sudoku is to fill the other empty cells with integers from 1 to 9, such that each number appears exactly once in each row, exactly once in each column, and exactly once in each region. The numbers given at the beginning ensure that the Sudoku puzzle has a unique solution. For example, the unique solution to the Sudoku in Figure 2.1 is illustrated in the right panel. We measure performance in the Sudoku task by the number of correctly filled cells.

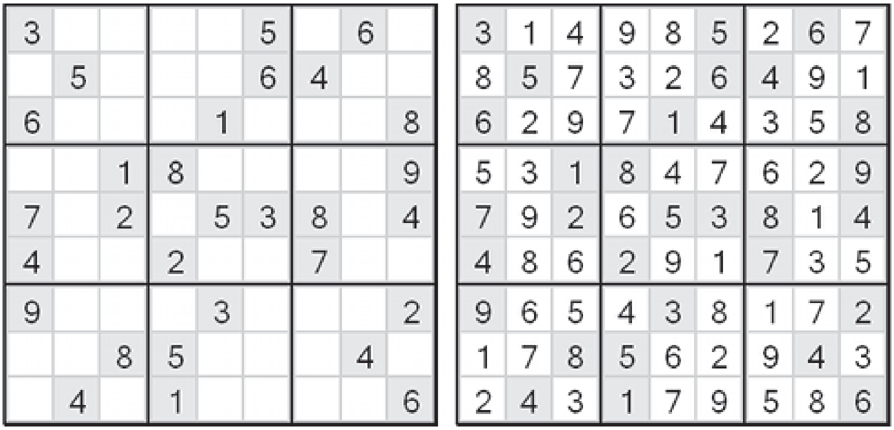
When solving a Sudoku puzzle, solutions often come in waves. Multitasking can be appealing when one is stuck: one can work on the other task and hope to see the problem from a different angle when switching back.

The other task was to find as many words as possible in a Word Search puzzle. An example of a Word Search puzzle is presented in the left panel of Figure 2.2, and its solution is presented in the right panel. Participants had to look for the English names of European and American countries in a 17x17 letter grid. Words could be in all directions, including diagonal and backwards. Subjects' performance is measured by the number of correct words found.[13]

[12] Despite being a numbers-based game, no mathematical operation is needed in Sudoku and men are not seen as better at Sudoku. Championships are always mixed and often won by women. Furthermore, there is no gender difference in Sudoku performance in our sample. In Round 1, women on average score 116 points and men 119 points (p=0.72; t-test).

[13] Subjects did not know in advance how many words were hidden in the puzzle, but they knew that

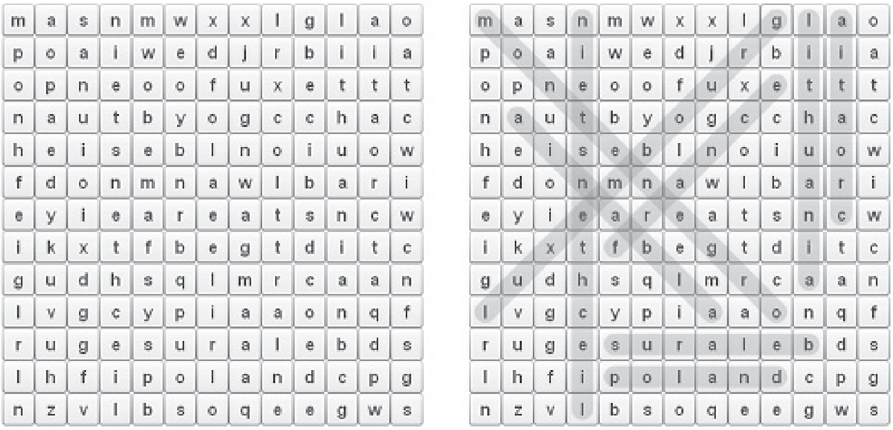
Figure 2.1. Sudoku



3					5		6	
	5				6	4		
6				1				8
		1	8					9
7		2		5	3	8		4
4			2			7		
9				3				2
		8	5				4	
	4		1					6

3	1	4	9	8	5	2	6	7
8	5	7	3	2	6	4	9	1
6	2	9	7	1	4	3	5	8
5	3	1	8	4	7	6	2	9
7	9	2	6	5	3	8	1	4
4	8	6	2	9	1	7	3	5
9	6	5	4	3	8	1	7	2
1	7	8	5	6	2	9	4	3
2	4	3	1	7	9	5	8	6

Figure 2.2. Word Search



m	a	s	n	m	w	x	x	l	g	l	a	o
p	o	a	i	w	e	d	j	r	b	i	i	a
o	p	n	e	o	o	f	u	x	e	t	t	t
n	a	u	t	b	y	o	g	c	c	h	a	c
h	e	i	s	e	b	l	n	o	i	u	o	w
f	d	o	n	m	n	a	w	l	b	a	r	i
e	y	i	e	a	r	e	a	t	s	n	c	w
i	k	x	t	f	b	e	g	t	d	i	t	c
g	u	d	h	s	q	l	m	r	c	a	a	n
l	v	g	c	y	p	i	a	a	o	n	q	f
r	u	g	e	s	u	r	a	l	e	b	d	s
l	h	f	i	p	o	l	a	n	d	c	p	g
n	z	v	l	b	s	o	q	e	e	g	w	s

m	a	s	n	m	w	x	x	l	g	l	a	o
p	o	a	i	w	e	d	j	r	b	i	i	a
o	p	n	e	o	o	f	u	x	e	t	t	t
n	a	u	t	b	y	o	g	c	c	h	a	c
h	e	i	s	e	b	l	n	o	i	u	o	w
f	d	o	n	m	n	a	w	l	b	a	r	i
e	y	i	e	a	r	e	a	t	s	n	c	w
i	k	x	t	f	b	e	g	t	d	i	t	c
g	u	d	h	s	q	l	m	r	c	a	a	n
l	v	g	c	y	p	i	a	a	o	n	q	f
r	u	g	e	s	u	r	a	l	e	b	d	s
l	h	f	i	p	o	l	a	n	d	c	p	g
n	z	v	l	b	s	o	q	e	e	g	w	s

As in the case of Sudoku, it is reasonable to expect subjects to switch when unable to find new words for a while. The situation is similar to polishing a paper, when reading the same lines over and over becomes counterproductive after a while – one changes to another task simply because a 'fresh eye' is needed to recognize meaning behind the letters.

2.3.3 Procedures, payments, timeline

One pilot and ten regular sessions were run in the computer lab of CREED (Center for Research in Experimental Economics and Political Decision-Making) at the University of Amsterdam. Participants were university students from various fields of study. The application procedure ensured that the two genders were represented approximately equally in every session, but left subjects unaware that the experiment examines gender-related issues. The experiment was conducted in English, therefore both international and Dutch students could participate. All instructions and tasks were computerized,[14] and subjects were not allowed to use any paper or take notes during the experiment.

The experiment started with an introduction that explained the rules of the two tasks and gave the participants opportunity to practice. Subjects learned that there would be two rounds and that they would have to play a Sudoku and a Word Search in both rounds. In each round, subjects earned 6 points for each correctly filled Sudoku cell and lost 6 points for each cell filled with a wrong number to avoid random guessing. Subjects were not penalized for cells filled with multiple numbers.[15] They received 9 points for each word found in Word Search. In Word Search, only entire words could be marked and there was therefore no need to penalize random clicking. Subjects' total points for each round were determined as the sum of their points in Sudoku and their points in Word Search. Negative total points were set to 0. One of the two rounds was randomly selected for payment at the end and the conversion rate was 1 euro per 11 points. In addition to this, there was a fixed show-up fee of 7 euros. The performance payments and the conversion rate were chosen based on the results of a pilot, such that subjects could earn approximately equal amounts on the two tasks and that the average payment was around 23 euros. The sessions lasted for approximately 1 hour and 45 minutes.

The order of the tasks within each round was randomized, and the assignment of subjects to the three treatments in round 2 was random as well, so that each group consisted of approximately one third of the subjects in every session. The rules of the

they would be notified once all words were found.

[14] The program was written in PHP (an HTML-embedded scripting language) and was displayed using the web browser Mozilla Firefox.

[15] Subjects could enter multiple numbers in one cell to denote uncertainty.

Table 2.2. Number of observations per cell

	Men	Women	Sample
Group 1	30	40	70
Group 2	39	31	70
Group 3	43	35	78
Total	112	106	218

treatments were explained immediately before the start of the treatment. Subjects were not aware of the fact that not everyone was playing the same treatment as they did.

After both rounds were over, but before being informed about their payment, we elicited some background information such as gender, age, field of study, and nationality from the subjects through a questionnaire. Those who participated in Treatment Choice were also asked their reasons for (not) switching.

2.3.4 Data

Our sample consists of 218 subjects from the ten regular sessions.[16] They are 22 years old on average and the majority of them is Dutch (73 percent). Approximately half of the sample consists of economics students (53 percent). The sample contains 11 censored observations from subjects who solved the entire Sudoku puzzle in the second round but not in the first.[17] As Section 2.3.1 explained, subjects were randomly assigned to three groups. Table 2.2 shows the number of observations per group and gender.[18] As we can see, there are between 30 and 43 subjects per cell.

[16] We only use the data from the regular sessions because some parameters were changed after the pilot session.

[17] In addition, 17 subjects solved the entire puzzle in the first round and 11 of these also in the second round. These 11 subjects are excluded since we do not know how their performance changed from the first to the second round. We also dropped the six subjects who solved the puzzle only in the first round. Otherwise we would encounter a sample selection problem: among the best performers of Round 1, we would only keep those who fall back enough in Round 2 to not solve the entire puzzle. Inclusion in the sample is thus conditional on not having solved the entire Sudoku in Round 1. Recall that every subject receives treatment Single in Round 1; therefore inclusion is independent of treatment.

[18] The distribution of the dropped subjects is as follows: 5 from Group 1, 8 from Group 2 and 4 from Group 3. Of the 17 dropped subjects, 14 are male and 3 are female.

2.4 Results

2.4.1 Multitasking and performance

Performance is measured as the sum of Sudoku plus-points and Word Search points.[19] Table 2.3 shows means per group and gender (for both rounds), and performance differences between rounds. Note that the difference-in-differences(-in-differences) strategy takes care of any performance differences between cells in Round 1. Results are qualitatively the same when using relative instead of absolute changes.

Comparing the results of Group 1 and Group 2 to each other shows that the productivity effects of multitasking are significantly negative: the difference-in-differences is -23 points (t-test: p=0.04). Subjects who could pick their own schedule (Group 3) perform only slightly better than those forced to multitask and score 21 points less than Group 1 (p=0.07).

The difference-in-differences in performance between men and women in Group 2 suggests that men handle forced multitasking relatively better than women, but the difference is not significant (p=0.62). The results of Group 3, on the other hand, suggest that women are better at organizing their own schedule, but this difference is not significant either (p=0.35). There are no gender differences in learning either: the performance improvement for Group 1 subjects is the same for both genders (p=0.84). In sum, a simple comparison of differences does not reveal any significant gender differences.[20]

Using regression techniques, we can check whether the results hold if we take censoring and the (non-significant) gender differences in learning into account. Table 2.4 shows the results of fixed effects and first-difference censored regressions which take full advantage of the panel structure of our data.[21] As we can see, the results of the censored regressions are very close to the results of the fixed effect estimates and all the previous conclusions are confirmed. The coefficients of Treatment Multi and Treatment Choice (relative to Treatment Single) are negative and significant at the 5 percent and the 10 percent level, respectively.[22] The gender-specific estimates confirm that there is no gender difference in learning (the gender dummy is insignificant). The point estimates suggest that men adapted better to Treatment Multi and women adapted better

[19] Sudoku minus-points were only used to discourage random guessing, not to measure performance. There is no significant gender difference in minus points in Round 1. The same is true for Round 2, where there is no significant gender difference in any of the treatments.

[20] Non-parametric ranksum tests lead to the same conclusions as the t-tests in all cases.

[21] Note that since there were two rounds, first-difference and fixed effects estimates are equivalent.

[22] Our results are primarily driven by Sudoku scores while there is little difference in Word Search scores between treatments. A possible explanation is that in Sudoku, solving the next cell depends crucially on information discovered while solving previous cells and interrupting the game to multitask can therefore really harm performance. It is also possible that for Word Search a negative effect of multitasking is outweighed by a positive effect of a 'fresh eye'.

Table 2.3. Average total points per cell

	Group 1 (n=70)			Group 2 (n=70)			Group 3 (n=78)		
	Round 1	Round 2	Diff.	Round 1	Round 2	Diff.	Round 1	Round 2	Diff.
Men	184	188	4	186	172	-14	195	174	-22
Women	185	192	7	198	177	-21	205	198	-7
Both	185	190	6	191	174	-17	200	185	-15

Note: all numbers are rounded to the nearest integer.

to Treatment Choice, but none of these gender differences is significant.[23]

Table 2.4. Impact of treatments on total points

	Group-specific estimates		Gender-specific estimates	
	FE	Censored	FE	Censored
Treatment Multi	-22.76**	-24.34**	-28.39**	-31.09**
	(10.98)	(11.44)	(12.98)	(13.75)
Multi×Male			10.90	13.37
			(21.78)	(22.49)
Treatment Choice	-20.97*	-21.19*	-14.14	-15.51
	(11.33)	(12.02)	(16.30)	(17.26)
Choice×Male			-11.63	-9.12
			(23.11)	(24.17)
Male			-3.40	-5.49
			(16.39)	(17.17)
Nr. of obs.	218	218	218	218

Note: Robust standard errors are shown in parentheses; significance levels: *** p<0.01, ** p<0.05, * p<0.1.

2.4.2 Propensity to multitask

To examine gender differences in the propensity to multitask, we use the results of Group 3.[24] Table 2.5 describes the switching behavior of men and women in Treatment Choice. As we can see, 71 percent of the subjects do actually switch when they are allowed to and the share of switchers is exactly the same for men and for women. So contrary to the claims of Fisher (1999), men do not focus on a single task any more than women do. Moreover, we can reject that women switch more or equally often than men (one-sided t-test; p-value=0.06).

Table 2.5. Number of switches in Treatment Choice

	Men	Women	All
Mean	2.50	1.74	2.16
Standard deviation	2.53	1.67	2.20
Share of switchers	0.71	0.71	0.71
Number of observations	42	35	77

Table 2.6 displays the results of two OLS regressions where the number of switches is the dependent variable. In Column 1, we only control for performance in Round

[23] Neither is the difference between the Multi×Male and Choice×Male coefficients significant (p=0.30; Wald test after FE regression).

[24] We excluded one subject from this analysis because he misused the 'Switch' button (switched multiple times within the same second).

1, while in Column 2 we include session and task-order fixed effects. Contrary to our expectations, performance in Round 1 does not influence switching behavior at all; this also implies that the impact of gender on switching is not caused by performance differences. When task order and session fixed effects are also included, the gender difference becomes significant at the 10 percent level. In sum, the results show that if there is any gender difference, it is men switching more than women and not the other way around.

Table 2.6. Regression results on propensity to switch

Dependent variable: nr. of switches	(1)	(2)
Male	0.76	0.92*
	(0.50)	(0.53)
Points in Round 1	0.00066	0.00536
	(0.0046)	(0.0050)
Nr. of obs.	77	77
Task order and session FE	no	yes

Note: Robust standard errors are shown in parentheses; significance levels: *** p<0.01, ** p<0.05, * p<0.1

It is interesting to look at how the performance drop in Group 3 relates to the number of switches. In Table 2.7, we regress the performance difference between Round 1 and Round 2 for subjects in Group 3 on the number of switches. The performance difference is insignificantly negatively correlated with the number of switches. When we restrict the sample to those who actually switch at least once in Column 2, the coefficient becomes significant at the 10-percent level. This indicates that the performance of subjects who switch more often suffers more. But we have to be careful in interpreting this coefficient. Although by using differences we take into account baseline performance levels, the number of switches might still be endogenous with respect to learning or tiredness effects.

So why do subjects switch although this seemingly harms their performance? Subjects already experienced an example of each task in Round 1 which should minimize switching due to mere curiosity. Indeed, the average subject (amongst those who switch) switches for the first time after 225 seconds. The second switch, for those who switch at least twice, on average occurs after another 237 seconds. Moreover in the post-experimental questionnaire, many subjects explicitly stated 'looking at the problem with a fresh eye' as a reason for switching while none mentioned curiosity. It therefore seems more likely that subjects switched because they (wrongly) thought it increases their performance.

Table 2.7. Performance in Treatment Choice and the propensity to switch

Dependent variable: performance change of Group 3 subjects	(1)	(2)
	All:	Switchers:
Nr. of switches	-3.97	-6.26*
	(3.21)	(3.46)
Nr. of obs.	77	55

Note: Robust standard errors are shown in parentheses; significance levels: *** p<0.01, ** p<0.05, * p<0.1.

2.5 Discussion and conclusions

Our results demonstrate that work schedules can be an important determinant of productivity. We find that multitasking significantly lowers performance as compared to a sequential execution. This suggests that the costs of switching, which include recalling the rules, details and steps executed thus far, outweigh the benefit of a 'fresh eye'.[25] Subjects who could choose the amount and timing of their switches freely did only marginally better than those forced to switch at unanticipated points in time and they perform significantly worse than those working under the exogenously imposed sequential schedule. Finally, we find no evidence that women are better at (or more attracted to) multitasking.

The finding that subjects are unable to organize their own work optimally is not unprecedented. For example, Ariely and Wertenbroch (2002) find that students who can set their own deadlines perform worse than those forced to adhere to equally spaced deadlines. Possibly, subjects pick a suboptimal schedule because the two tasks imply a high cognitive load that leads to more impulsive choices, as suggested by the results of Shiv and Fedorikhin (1999).[26] Another possibility is that even though subjects choose the best schedule possible, planning itself requires so much effort that their performance on the tasks takes a hit.

The fact that in our experiment the number of switches is negatively correlated with performance supports the interpretation that subjects choose a suboptimal schedule. The hypothesis that the effort required for planning when to switch is at the root of the performance impact is however supported by the fact that the average number of switches in Treatment Choice is only 2.16, but subjects still fall back almost as much

[25] Subjects clearly do expect a benefit from a 'fresh eye'. Of those who switched, many explicitly stated 'looking at the problem with a fresh eye' as the main reason.

[26] The underlying idea is that when mental processing resources are depleted, spontaneous affective reactions tend to determine choice. Shiv and Fedorikhin (1999) find that subjects are more likely to pick an unhealthy food option over a healthy one when simultaneously trying to remember a seven-digit number. See Fudenberg and Levine (2006) for a theoretical model which can account for such effects.

as subjects in Treatment Multi who were forced to switch four times and could not anticipate the timing of the switches. It is difficult to distinguish between these explanations as the number of switches is potentially endogenous to performance. Whichever explanation is correct, the results are not in favor of self-imposed work schedules.

The results support the intuition that scheduling is an important input in the production function that deserves more attention in the economic literature. However, there are some caveats which need to be taken into account when extrapolating our findings. The results were obtained in a stylized lab setting and may be specific to the chosen tasks. Note also that we compared multitasking to sequential execution keeping time allocation between tasks constant, so our analysis does not extend to situations where time allocation varies. Future experiments could uncover whether individuals are able to optimally allocate their time between multiple tasks and whether there are gender differences in this regard. Furthermore, our strict time constraints and performance-dependent pay scheme possibly put pressure on the subjects which may affect performance.[27] This does not affect the internal validity of our results, as these factors are constant across treatments, but may mean that our results are particularly applicable to high pressure work situations. Furthermore, some potential benefits of multitasking are eliminated in our design. Multitasking may make work more stimulating which in turn could increase productivity. Workers could even repay a less boring work design with increased productivity as a form of gift exchange (Akerlof, 1982). The 12 minutes subjects work on a task in our experiment might be too short to get bored and there is no room for gift exchange. Further research is therefore needed to determine to what extent our results carry over to specific work environments.

If they do carry over, there are important implications for job design. Although our experiment does not provide a direct test of this, the results suggest that assigning multiple tasks to a worker may be problematic for reasons different from those suggested by the previous literature (e.g. by Holmstrom and Milgrom, 1991). Namely, if workers are given several tasks at once, they may hamper their own productivity by juggling between the tasks. One way to avoid this problem is to assign the next task only after the previous one has been finished. Another way is to prescribe a sequential execution rather than letting workers choose their own schedule.

The finding that subjects perform worse under the self-chosen work schedule also adds a new aspect to the debate about the centralization of decision making. The standard argument in favor of decentralization is that workers have more information than managers and that more decision making rights lead to an increase in motivation.[28]

[27] The experiments following Duncker's candle problem (Duncker, 1945) show that a 'high drive' (such as a payment scheme that depends on time) can affect performance (see Glucksberg, 1962 and Spence et al., 1956).

[28] For example, Falk and Kosfeld (2006) find that agents perceive principals' controlling decisions as

Typically, loss of control is mentioned as the sole disadvantage. Our results suggest further issues: decision-making may take away resources from a worker's actual tasks and workers may simply not be able to schedule their own work optimally. One limitation of our study though in this regard is that there is little room for learning in our experiment. Over time, workers may get better at choosing their own schedules or learn to avoid multitasking. Future research could uncover whether with more experience or a longer time horizon subjects are able to optimize their schedule.

As far as gender differences are concerned, we do not find any evidence for them in the effects of multitasking. Besides, the share of switchers is exactly the same for men and women and the average number of switches is higher for men. These results contradict the claims of Fisher (1999): if men think so much more linearly than women, why don't they insist more on a sequential schedule? And why is it that women do not adapt better to multitasking than men when forced to alternate? In sum, the view that women are better at multitasking is not supported by our findings.

a sign of distrust and reduce their performance when being controlled.

Chapter 3

The effect of a sibling's gender on earnings, education and family formation[1]

3.1 Introduction

Family environment has long been considered to be a determinant of children's long term outcomes (Haveman and Wolfe, 1995). The influence of close family members is of particular interest since they play a significant role in most people's lives. This applies among others to siblings. According to time use surveys, children spend more out-of-school time with their siblings than with anyone else (McHale and Crouter, 1996). Research on sibling relationships shows that siblings provide reference points to each other, both in childhood and in adulthood (Adams, 1999; Keim et al., 2009). While these observations suggest that siblings have a big potential to have an impact, evidence on causal effects in sibling relationships is scarce (see Joensen and Nielsen, 2015; Altonji et al., 2013; Nicoletti and Rabe, 2014 and Dahl et al., 2014 for exceptions).[2] We try to shed more light on this by studying the role of one particular factor: we examine how the gender of a sibling affects individual's education, earnings and family formation.

The gender of the sibling is interesting because having a brother implies a different family environment than having a sister. Several studies show that men and women differ in many regards, which means that brothers and sisters are different too. For example, women are less competitive, more risk averse, less likely to negotiate and more socially minded than men (Croson and Gneezy, 2009; Bertrand, 2010). As compared to men, women also marry and have children at a younger age, are less likely to get

[1] This chapter is based on Peter et al. (2015).

[2] In contrast, studies on the influence of parents are numerous (see Holmlund et al., 2011 for a review).

promotions and earn less (Copen et al., 2012; Baizán et al., 2003; Bertrand, 2010). Moreover, men and women are also treated differently by parents, both in childhood and in adulthood. For example, parents encourage different types of play and buy different types of toys for sons than for daughters (Fisher-Thompson, 1993; Nelson, 2005). In adulthood, parents support the family formation of daughters more than the family formation of sons, for example by providing more informal childcare (Danielsbacka et al., 2011; Pollet et al., 2009). In sum, those with a brother are exposed to different family conditions than those with a sister. We examine whether this has an impact on their labor market outcomes and family formation; we focus on these outcome variables because many gender differences relate to these domains.

The first studies on the role of siblings' gender were done by psychologists who wanted to know how it affects child development.[3] The focus has been on the impact of older siblings' gender because older siblings are typically more dominant in sibling relationships (Tucker et al., 2010). The results of this literature indicate spillover effects, as they show that children with (older) sisters are more feminine/less masculine than children with (older) brothers. Economists examined whether there is an impact on educational outcomes. None of the studies found an impact on white men, while results on white women were controversial.[4] Butcher and Case (1994) found that women with any sisters attained lower education than women with only brothers. Their preferred explanation was that the presence of a second daughter changes the reference group for the first: the girls are grouped together and get lower educational standards than boys. However, their results on women could not be replicated by Kaestner (1997) and Hauser and Kuo (1998).[5]

These mixed findings make it difficult to draw conclusions from the previous literature. Further, the above studies typically control for family size, treating it as a fixed variable. However, this assumption was questioned in the seminal paper of Angrist and Evans (1998), which shows that family size depends on children's sex composition. This is because parents may prefer certain sex compositions over others, so the gender of the children affects subsequent fertility decisions. This has two important consequences for empirical identification. First, family size is an outcome variable and therefore controlling for it leads to "bad control bias" (Angrist and Pischke, 2009). Second, a selection bias can arise, irrespective of whether family size is controlled for or not. This is because parental preferences imply that the gender of an earlier born child influences the selection of a potential later born child into the sample. As we show

[3] The pioneers were Koch (1955) and Brim (1958). See Rust et al. (2000) for a review and further analysis.

[4] We focus on results on whites from Western countries. For results on other populations, see e.g. Parish and Willis (1993), Morduch (2000), Chen et al. (2013) and Jayachandran and Pande (2015).

[5] Similarly, a mimeo by Pettersson-Lidbom et al. (2008) finds little evidence that child gender has an effect on siblings and parents.

in Section 3.2, this can lead to biased estimates since parental preferences may affect children's outcomes as well.

We apply an empirical approach that circumvents these problems. Using a sample of dizygotic (i.e. non-identical) twins from Sweden, we compare men (women) with co-twin brothers to men (women) with co-twin sisters.[6] We examine whether the gender of the co-twin has an impact on their education, earnings and family formation. Our identification strategy exploits the fact that twins are born at the same time. This implies that there is no selection bias in this setup: parents cannot make decisions about one twin based on the gender of the other twin, given that the twins are born only minutes apart.[7] Family size can still be affected, because parents can choose the number of additional children. We avoid the bad control problem because we do not control for family size but examine it separately as a potential channel.

It is important that we can distinguish the twins by zygosity. While the sex of dizygotic co-twins is random, the sex of monozygotic twins is always the same as the sex of their co-twin (see Section 3.2). Therefore estimates on the sample of all twins could suffer from "zygosity bias"; that is, the coefficient of the co-twin's gender could pick up potential differences between dizygotic and monozygotic twins. Our data contains high quality information on zygosity, so we avoid this problem by restricting the main estimation sample to dizygotic twins. In addition, we can empirically assess the magnitude of the zygosity bias by comparing estimates on dizygotic twins to estimates on the full sample of twins. We find that zygosity bias can lead to misleading conclusions primarily when the outcomes of women are examined.

While economists have long used twin samples, our approach is different from what is typical in the literature. Most researchers apply twin fixed effects, as they want to exploit that monozygotic twins are very similar to each other. We point out that twins have another advantage. Since they are born at the same time, we can use them to avoid the above-mentioned selection bias. Thus, we exploit twins in a non-traditional way (see also Gielen et al., 2016).

A co-twin's gender can have an impact in various ways. Effects can arise via childhood spillovers, reference point considerations or changes in parental treatment - we will refer to all of these processes collectively as "social mechanisms". In addition, there could be an impact via the above-mentioned family size channel as well. Finally, our twin design gives rise to the possibility of a biological channel, which could occur if hormonal transfer between twins was possible (Miller, 1994). We discuss these three potential mechanisms in detail and investigate which one of them is leading the results in Section 3.5.

[6] We follow the previous literature and analyze men and women separately.

[7] Selective abortion and IVF techniques were not available in the time period that we examine.

We find that the gender of the co-twin influences both men and women, but in a different way. Men with brothers earn more and are more likely to get married and have children. In case of women, there is an impact on education and age at first birth: women with sisters obtain lower education and give birth earlier. Our analysis shows that it would be difficult to explain the results with the family size channel or with hormonal transfer. Instead, the most likely explanation is that social mechanisms are at work. These mechanisms can operate in case of singletons as well, although their effect might be weaker, especially as far as the influence of a younger or more widely spaced sibling is concerned.

We contribute to the literature in several ways. We look at a wider range of outcomes than previous studies. We also point out the methodological challenges of empirical identification and offer a solution. As a result, we provide new insights into the impact of siblings' gender. In addition, our study has important implications for research on other family factors. In particular, children's sex composition is often used as an instrument to identify the impact of family size. We point out several factors that can question the validity of this instrument and therefore we suggest to use alternative approaches instead (see Section 3.5.1).

The remainder of this chapter unfolds as follows. Section 3.2 explains our empirical approach in more detail and Section 3.3 describes the data. We present the results in Section 3.4 and discuss potential explanations in Section 3.5. Finally, Section 3.6 concludes.

3.2 Empirical strategy

Several studies show that the gender of current children influences parity progression, that is, the probability of having additional children (see Angrist and Evans, 1998; Conley and Glauber, 2006; Åslund and Grönqvist, 2007; Dahl and Moretti, 2008; Cools and Kaldager Hart, 2015). In developed countries, the most common finding is that parents of two boys or two girls are more likely to have additional children than parents of a boy and a girl. Clearly, this indicates that there are parental preferences for children's sex composition. At the same time, the exact nature of these preferences is less obvious.

For example, the above phenomenon might arise because some families have a preference for gender mix. However, another possibility is that nobody prefers a gender mix, but some families have a preference for boys while others have a preference for girls. According to this explanation, some parents with same-sex kids will proceed to a next child because their first two children are not from the preferred sex. Families with a boy and a girl will not proceed, because they have a child from the preferred gender for sure. Note that for this explanation to be true, it is not necessary to observe differential progression after the first child. If the number of boy-preferring families equals the number of girl-preferring families, the same number of families will proceed to a second child after a first-born boy than after a first-born girl.

As this example demonstrates, parental preferences can be of various types and different preferences may cancel each other out in aggregate figures. Therefore parental preferences can always be present, even when there is no difference in parity progression ratios.[8] This means that we should always keep in mind that comparing individuals with an older brother to individuals with an older sister can lead to biased results. This is because families that proceed to a next child after having a boy may have different preferences than families that proceed after a girl. If people with different preferences raise their children differently, the estimates will be biased.

This intuition corresponds to a selection bias problem, as in Heckman (1979). To show this in detail we need to consider a set of equations. We will analyze men and women separately so parameters can be different by gender. Nonetheless, for the ease of exposition we omit subscripts for own gender. We start with the equation for the latent variable Y_i^*:

$$Y_i^* = \alpha + \beta^o G_i^o + \gamma X_i + U_i \tag{3.1}$$

where Y_i^* is the outcome of individual i, G_i^o is the gender of the older sibling and X_i denotes observable exogenous covariates. U_i consists of other relevant variables with $E\left[U_i\right] = 0$. Parental preferences are denoted by P_i and they are included in U_i. They are relevant because parents with different preferences may raise their children differently. They are unobservable, so they have to be in U_i instead of X_i. Nonetheless, this does not lead to bias in the estimation of equation (3.1) because G_i^o and P_i are independent, due to the random assignment of gender.[9]

The problem is that we cannot estimate equation (3.1) because Y_i^* is a latent variable. Instead of observing Y_i^*, we observe

[8] Parity progression ratios provide one-sided information: they can be used to show that some kind of parental preferences are present, but they cannot be used to prove the absence of all kind of preferences. It is easy to create examples where there is no differential progression at all but parental preferences are still present. For example, suppose that the same number of families prefer boys and girls. In addition, suppose that these families want to have *at most* two kids. It is clear that in this case, there will be no differential progression, neither at parity 1, nor at parity 2, even though parental preferences are present.

[9] Medical technologies for sex selection were unavailable in the time period that we examine. As far as natural sex selection is concerned, Wilcox et al. (1995) found that the timing of intercourse in relation to ovulation has no significant effect. Similarly, Gray et al. (1998) found no evidence for the hypothesis that maternal hormones influence sex selection. All in all, the established view among biologists is that sex is essentially random (Reece et al., 2010, p. 290).

$$Y_i = \begin{cases} Y_i^* & if \quad S_i = 1 \\ missing & if \quad S_i = 0 \end{cases} \tag{3.2}$$

That is, we observe outcomes for a selected sample: only for those people who were actually born ($S_i = 1$). This selection depends on both parental preferences (P_i) and the gender of the older child:

$$S_i = f(P_i, G_i^o) \tag{3.3}$$

If we try to estimate the parameters using the selected sample, we need to consider the expected value of Y_i conditional on $S_i = 1$:

$$E\left[Y_i | S_i = 1, G_i^o, X_i\right] = \alpha + \beta^o G_i^o + \gamma X_i + E\left[U_i | S_i = 1, G_i^o, X_i\right] =$$

$$\alpha + \beta^o G_i^o + \gamma X_i + E\left[U_i | f(P_i, G_i^o) = 1, G_i^o, X_i\right] \tag{3.4}$$

The last term makes the selection bias visible. We condition on both $S_i = 1$ and G_i^o. Since selection depends on both P_i and G_i^o, this implicitly defines P_i: only certain parents will proceed to a next child after they have a girl (boy). Recall that U_i includes P_i because parental preferences may affect how children are raised. This implies that $E\left[U_i | f(P_i, G_i^o) = 1, G_i^o = 1, X_i\right] \neq E\left[U_i | f(P_i, G_i^o) = 1, G_i^o = 0, X_i\right]$, so the estimates suffer from selection bias.

We circumvent this problem by using a sample of dizygotic (DZ) twins. Since twins are born at the same time, parents cannot make decisions about one twin based on the sex of the other twin. This means that there is no selection bias in this setup. Thus, we can estimate the following equation, for men and women separately:

$$Y_i = \alpha + \beta^{co} G_i^{co} + \gamma X_i + U_i \tag{3.5}$$

where Y_i denotes the outcome of twin i, G_i^{co}denotes the gender of the co-twin and X_i denotes exogenous covariates (birth year fixed effects). We will estimate standard errors such that we allow clustering at the level of the family of origin.

We do not control for family size because that could lead to bad control bias. In principle this means that the estimates have to be interpreted as a total impact that includes the family size channel. Nonetheless, our setting has two important features that make the family size channel less likely to act. First, twins are not necessarily the first children in the family. Therefore their sex might not be so influential, given that family size depends on the sex composition of all existing children. Second, twins already increase family size substantially, so we can expect their sex to have little additional impact. Indeed, in Section 3.5.1 we analyze family size separately as an outcome and find that the gender of the co-twin has no significant effect on it in our sample. This means that our estimates are not driven by the family size channel. Note that with a different identification strategy this might be more difficult to assess. For example, suppose that one were to analyze whether the gender of younger children affects the outcomes of older children. This empirical approach would also avoid the selection bias, since the gender of younger children cannot affect the existence of older children retroactively. However, in that case it could be more ambiguous whether the estimates reflect anything more than the family size channel.

We estimate equation (3.5) under the assumption that the gender of the co-twin is random. This is why we focus on dizygotics only. Just like two singleton siblings, DZ twins result from the separate fertilization of two different ova by two different sperms (Phillips, 1993). Their sex is determined the same way as the sex of singletons: the offspring will be male (female) if the sperm bears a Y (X) chromosome. Since each ovum is fertilized by a different sperm, the sex of a DZ co-twin is just as random as the sex of a singleton sibling. Thus, the probabilities of dizygotic males and females are thought to be approximately equal, independently of the sex of the other twin.[10] Indeed, the share of same-sex co-twins is fifty percent in our sample of dizygotic twins, for both men and women.

The case of monozygotic (MZ) twins is different because they arise when one zygote splits into two genetically identical units. This implies that MZ twins are all same-sex. Thus, if we did not differentiate by zygosity, differences between opposite-sex and same-sex twins could reflect not only the impact of the sibling's gender, but also the impact of having a genetically identical sibling or other potential differences between monozygotic and dizygotic twins. We circumvent this problem by restricting the main estimation sample to DZ twins.

3.3 Data

Our data comes from the SALT project (Screening Across the Lifespan Twin Study) of the Swedish Twin Registry (STR) (see Lichtenstein et al., 2002). The aim of the SALT was to survey all Swedish twins born in 1926-1958, irrespective of the sex composition

[10] This view is so widely accepted that researchers typically calculate the number of dizygotics within the twin population as twice the number of opposite-sex twins (Benirschke et al., 2012). This is the so-called Weinberg's differential rule (WDR). Although James (1979) raised questions about this method, several studies found that it is supported empirically (Vlietinck et al., 1988; Husby et al., 1991; Fellman and Eriksson, 2006). Thus, Elston et al. (2002) conclude that this is a reasonable approximation and actual deviations are small.

of the twin pair.[11] For these cohorts, the STR had records of every twin birth from the national birth register.[12] They contacted all available twins for a phone interview. The process started with a pilot in 1996-1997 and then full scale data collection took place in 1998-2002. The oldest cohorts were surveyed first, followed by younger cohorts.[13] The interviewers tried to get as many responses as possible; e.g. people were called back at least five times if they were unavailable.

In short, the SALT was a systematic screening of all available twins. Unfortunately, we do not have access to the universe of twins, so we cannot compare respondents to the rest of the twins. However, we know that the response rate was remarkably high: 74 percent. Thus, the STR managed to survey a large sample of the underlying population. Recall that the use of a survey is essential for our purposes: otherwise, we could not differentiate between MZ and DZ twins.

The SALT determines zygosity based on questions about intrapair similarity in childhood. This classification was validated in the pilot: 13 DNA-markers were analyzed in blood samples. Zygosity assignment proved correct in 99% of the cases (see Lichtenstein et al., 2002).

The survey provided information on marital status, children and siblings as well. In case of marriage, we use the dummy "ever married" that takes one if the individual has ever been married or cohabited and zero otherwise. The questions on children and siblings refer to biological children/siblings (other than the co-twin) who are still alive. We look at the number of children to analyze fertility.[14] To analyze timing, we calculate age at first birth from the birth year of children. In addition, we also ac-

[11] The SALT sample contains twins from cohorts 1886-1925 as well. However, these cohorts were screened only partially; many twins were left out because they did not take part in previous surveys. The early surveys targeted only certain sex compositions (typically same-sex pairs). In turn, the SALT sample contains disproportionately few opposite-sex twins from cohorts 1886-1925. We do not use these early cohorts in our research, only the complete cohorts of 1926-1958.

[12] This is remarkable as many twin registries have to enroll twins in alternative ways instead. For example, at the start of this project we obtained data from the Australian Twin Registry (ATR). However, it turned out that the ATR used specific advertisements for recruitment so that the sample composition of same-sex and opposite-sex twins was different. After discovering this problem, we abandoned the Australian sample. Nonetheless, it is reassuring that the preliminary results on that sample were quite similar to the results obtained here.

[13] Of course "younger" is meant in relative terms; participants are 41-74 years old. The STR focused on these ages because they wanted to screen for some diseases. A population above 40 is actually quite fortunate for our purposes, since fertility is typically completed by that age. It could have been a disadvantage for the analysis of income, but as we explain soon, we collected information on income from earlier years as well. Thus, we can look at income in the working ages.

[14] We imputed data for childless women. This is because women who never gave birth got 'missing' assigned for the children variables, even though their number of children should be zero. We do not have access to the variable on childbirth so we do not know which observations are truly missing and which ones should be recorded as zero. However, the questions on children follow directly the ones on siblings, so we inferred who could have answered them. In particular, a woman is assumed to have zero children if she answered the question on siblings but not the question on children. Since data on men is complete, we can check the accuracy of our strategy. The results are reassuring: 99.7% of men who answered the question on siblings also answered the question on children.

quired information on age at menarche (first menstruation) from the survey, to address whether fertility patterns are led by a biological channel.

For the analysis of education and income we obtained registry data. Less than 0.1% of the subjects were dropped because their education and income data could not be found in the registers. The source of the education data is the education register (utbildningsregistret, UREG). Years of education is calculated from obtained degree. The income data comes from the national register on taxable income (IoT, Inkomst- och taxeringsregistret). This is *not* based on individuals' self-reports but on information collected from the employers and other agencies providing taxable benefits. We use records from years 1971-2007 (all converted to 2010 prices).[15] The income variable measures total taxable labor income in the given year: it includes wages, income from self-employment, pensions, sickness benefits and other taxable benefits.[16] From any given year, we use the available income records of those who are aged between 25 and 64 in that year. The main source of income in this age range is typically the labor market.

Since we have income records from many years, we have several observations per person. We take their average to get a permanent income measure for each individual. We use the log of this measure, ln(permanent income) as outcome variable. Note that people may have zero income in one year or another; however, almost nobody has zero income in all years. Therefore we lose hardly any observations when we take the log of permanent income.[17] On the other hand, we averaged over different number of observations in different cohorts. This implies potential heteroskedasticity, even though cohort fixed effects are included in the regressions. We will use robust standard errors to take this into account. In addition, later we will also show results from a pooled regression on ln(income). In this case, we pool the different years instead of taking the average, so more observations fall out due to zero income.[18] We run a pooled regression on the extensive margin as well. That is, we analyze whether there is an impact on the probability of having positive income.

Descriptive statistics are shown in Table 3.1. The first column shows the characteristics of DZ twins, our main estimation sample. The second and third column breaks

[15] In fact, the income data starts in 1968. However, after inspecting the data, we had concerns about sample selection in the first 3 years. In 1968-1970, 16.24% of the observations are missing, whereas data is almost complete after that period. Therefore we decided to use only the 1971-2007 data in the analysis.

[16] Many benefits became taxable in 1974 and therefore they are included in the income measure since then. This means that our variable was somewhat more restrictive in the pre-1974 years. Nonetheless, recent work by Björklund et al. (2009) suggests that this is not a major concern, as benefits constitute a very small share of total earnings. In any case, our results remain qualitatively the same if we restrict income data to 1974-2007.

[17] We lose only 0.08% of women. We do not lose any men.

[18] This share is still quite low, since the income measure includes all kinds of benefits. It is 4.46% in case of women and 0.76% in case of men.

down the sample by the gender of the co-twin. OS DZ denotes dizygotic twins whose co-twin is of the opposite sex, while SS DZ denotes dizygotic twins whose co-twin is of the same sex. Recall that the SALT surveyed individuals, not twin pairs, so inclusion in the sample is not conditional on the participation of the co-twin. In line with the previous discussion, we can see that the share of same-sex co-twins is approximately fifty percent, among both men and women. [19]

[19] There are somewhat more women in the sample, as they are more likely to respond to surveys (see Singer et al., 2000). Note that we analyze men and women separately, so this will simply mean that the analysis of the latter will be based on more observations.

Table 3.1. Descriptive statistics

	All DZ twins	OS DZ twins	SS DZ twins
Panel A: Men			
Same-sex co-twin	0.50	0	1
	(0.50)	(0)	(0)
Ever married	0.88	0.87	0.89
	(0.32)	(0.33)	(0.32)
Any kids	0.82	0.82	0.83
	(0.38)	(0.39)	(0.38)
Number of kids	1.89	1.86	1.91
	(1.24)	(1.24)	(1.23)
Age at first birth	27.87	27.88	27.85
	(5.19)	(5.15)	(5.23)
Number of siblings	2.00	2.01	1.99
	(1.85)	(1.89)	(1.81)
Years of education	11.11	11.11	11.11
	(2.67)	(2.66)	(2.68)
Ln (permanent income)	12.44	12.43	12.44
	(0.39)	(0.40)	(0.39)
Number of observations	13664	6890	6774
Panel B: Women			
Same-sex co-twin	0.50	0	1
	(0.50)	(0)	(0)
Ever married	0.90	0.90	0.90
	(0.31)	(0.31)	(0.31)
Any kids	0.87	0.88	0.87
	(0.33)	(0.33)	(0.34)
Number of kids	1.99	1.98	2.00
	(1.18)	(1.17)	(1.18)
Age at first birth	24.94	25.07	24.82
	(4.75)	(4.77)	(4.72)
Age at menarche	13.45	13.42	13.47
	(1.61)	(1.65)	(1.56)
Number of siblings	2.04	2.05	2.03
	(1.86)	(1.87)	(1.86)
Years of education	11.06	11.12	10.99
	(2.54)	(2.56)	(2.53)
Ln (permanent income)	11.92	11.93	11.91
	(0.59)	(0.57)	(0.60)
Number of observations	14950	7522	7428

Note: Marriage includes cohabitation. Standard deviations in parentheses.

3.4 Results

In this section we show results on the total impact of the co-twin's gender. These estimates can be the result of several mechanisms; we will discuss them after the main findings are presented.

The main results are reported in Table 3.2. For men, having a brother instead of a sister has no impact on years of education. This is consistent with the findings of previous literature. On income we see a positive impact that is significant at the 10 percent level. The gender of the co-twin also affects the probability of family formation: those with brothers are more likely to get married and have children. As a result, they have more children.[20]

The results on women show a different pattern. There is an impact on education: having a sister instead of a brother decreases years of education. This is in line with the findings of Butcher and Case (1994). The coefficient on income is also negative, but it is not statistically significant. There is no impact on the probability of family formation, but there is one on timing: age at first birth is lower for those with sisters. Thus, women have children earlier and obtain lower education if they have a sister instead of a brother.

Before going further we examine whether the results are sensitive to alternative specifications or definitions. First we check what happens if we use a probit or logit model to analyze the impact on the probability of getting married and on the probability of having kids. We find that our conclusions remain unchanged (untabulated results). Next we conduct some additional robustness checks, which are presented in Table 3.3. For education, we analyze whether the gender of the co-twin affects the probability of having more than primary education (that is, more than 9 years of schooling). This is an interesting threshold to look at because whenever researchers see a decrease in years of schooling and a decrease in age at first birth, a natural question is whether the two phenomena are related. A usual suspect for such a pattern is drop-out from high school due to childbirth. Therefore it is worthwhile to examine whether the impact on education is present already at the primary school stage, when childbirth is very uncommon. In case of income, we run pooled regressions on ln(income) and on the probability of having positive income. In these regressions we have several observations for every individual. Although this increases sample size, the standard errors do not shrink because of the clustering. To avoid the overrepresentation of those who could be observed for more years than others, we weight in these pooled regressions with the

[20] There is no significant difference in the intensive margin, that is, in the number of children for those who have at least one child. However, this estimate would be difficult to interpret as this sample is conditional on a variable that is significantly affected (having any children). For the same reason, we do not interpret the estimates on age at first birth.

Table 3.2. Estimates of the effect of a same-sex co-twin on education, income and family formation, on the sample of DZ twins

	Years of education	Ln(perm. income)	Ever married	Any kids	Number of kids	Age at first birth
Panel A: Men						
Same-sex co-twin	-0.002	0.013*	0.013**	0.013**	0.044**	-0.029
	(0.048)	(0.007)	(0.006)	(0.007)	(0.022)	(0.101)
N	13664	13664	13560	13434	13430	10979
Panel B: Women						
Same-sex co-twin	-0.098**	-0.011	-0.001	-0.005	0.013	-0.211**
	(0.043)	(0.010)	(0.005)	(0.006)	(0.020)	(0.087)
N	14950	14938	14873	14756	14756	12846

Note: The standard errors shown in parentheses are robust and allow clustering by the family of origin. All regressions include a constant and cohort fixed effects. Education is measured as years of schooling. Permanent income is average income in years 1971-2007 (if subject is aged 25-64 in the given year). The variable "ever married" takes one if the individual has ever been married or cohabited and zero otherwise. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

inverse of the number of observations. In this way we ensure that every individual is represented equally.[21] The pooled regressions also include year fixed effects.

In case of the family formation variables, we change the definition of the outcome variables. So far we examined whether the respondent has ever been married, but now we examine the probability of being married at the time of the interview. We redefine the variables "any kids" and "number of kids" such that they count only if the respondent lists the birth date of the kids as well. As far as age at first birth is concerned, we check whether the results are led by a few outliers whose first kid was born at an unusually early or late age. That is, we drop those whose age at first birth was less than 16 or more than 40. Setting the lower threshold to the age of 16 is also motivated by the fact that primary school lasts until that age.

Table 3.3 shows the results of these additional analyses. For men, the pooled regression on ln(income) shows the same positive effect of brothers as the previous estimate on ln(permanent income). There is no impact on the extensive margin, that is, on the probability of having positive income. The results on family formation are also very similar to Table 3.2.

The results on women are also in line with previous estimates. Those with a sister are less likely to have more than primary education, and the impact on age at first birth is actually slightly higher than before. These results suggest that the impact on education is not led by the impact on age at first birth. Further analysis shows that the effect on age at first birth arises because women with a same-sex co-twin have children earlier in their twenties than women with an opposite-sex co-twin (untabulated results).

For the sake of comparison, it could be interesting to see how the results would look like if we could not differentiate the twins by zygosity. Therefore we report results on the sample of all twins (including MZs) in the Appendix, in Table 3.5. For men, the results on income are essentially the same as in Table 3.2. The estimates on the family formation variables are also in the same direction, but the coefficients are somewhat smaller and the effect on the probability of having any kids is not significant any more. In case of women, the results in Table 3.5 are very different from the results in Table 3.2. The coefficients are insignificant and very close to zero for all outcomes, including education and age at first birth. Thus, the estimates on women would be quite misleading if we had no information on zygosity.

[21] Nonetheless, the results are qualitatively the same in unweighted regressions, where each observation is represented equally, not each individual.

Table 3.3. Robustness of the estimates of the effect of a same-sex co-twin on education, income and family formation, on the sample of DZ twins

	More than primary education	Probability of positive income	ln(income)	Married	Any kids	Number of kids	Age at first birth
Panel A: Men							
Same-sex co-twin	-0.006	0.001	0.013*	0.018***	0.013*	0.044**	-0.124
	(0.008)	(0.001)	(0.007)	(0.007)	(0.007)	(0.022)	(0.090)
N	13664	421434	418217	13560	13352	13352	10700
Panel B: Women							
Same-sex co-twin	-0.018**	-0.002	-0.007	0.000	-0.004	0.013	-0.227***
	(0.008)	(0.002)	(0.008)	(0.007)	(0.006)	(0.020)	(0.085)
N	14950	461791	441173	14873	14720	14720	12795

Note: The standard errors shown in parentheses are robust and allow clustering by the family of origin. All regressions include a constant and cohort fixed effects. More than primary education means more than 9 years of schooling. The second and third columns show pooled regressions on the probability of having positive income and on ln(income). These pooled regressions also include year fixed effects and are weighted such that every person is represented equally. Married takes one if the individual is currently married or cohabiting and takes zero for those who have never been married or whose marriage ended. Any kids and number of kids are counted only if the respondent lists the birth date of the kids as well. Age at first birth is counted only if it is between 16 and 40. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

3.5 Potential explanations

What drives the results of the previous section? There are several potential explanations that we need to discuss before we can answer this question. First, recall from Section 3.2 that the estimates have to be interpreted as a total impact that includes the family size channel. Thus, in the first subsection we will analyze this channel separately. In the second subsection we investigate another potential channel, called twin testosterone transfer. After the examination of these two alternative explanations, we discuss how the results can arise via social mechanisms such as childhood spillovers, reference point considerations and changes in parental treatment.

3.5.1 Family size

In this section, we focus on the family size channel as a potential explanation for the results. This channel is relevant only if two conditions are met: 1) the sex composition of the twins affects family size and 2) family size affects the outcomes of the twins. We can examine the first condition empirically. As the first column of Table 3.4 shows, there is no impact on the number of siblings. This is in line with our expectations: since twins already increase family size substantially, the sex composition of the pair does not increase family size further. In addition, parity progression is influenced by the sex composition of all existing children, while twins are typically not the first birth. Indeed, more than 60 percent of our sample has at least one older sibling. Of course one can argue that our measure is imprecise, since only those siblings are listed who are still alive at the time of the interview. However, this means that the share of those with older siblings is potentially even higher, which makes our argument even stronger. Nonetheless, we investigated whether the results change if we drop the oldest cohorts, since the survival problem should be mitigated in this case. Our conclusions remained the same.[22] Thus, the data does not support the family size explanation.

The results in Table 3.4 are one way to address the family size channel. Another way is to consider the second condition, that is, to examine whether we can expect family size to have an impact on labor market outcomes and family formation. In this respect it is very informative to look at Scandinavian studies that use twin births as instruments for family size. In particular, Black et al. (2005) find that education, earnings and teen births are all unaffected by family size in Norway. Åslund and Grönqvist (2010) reach similar conclusions about years of schooling and labor market outcomes in Sweden. This shows that even if we had no information on the number of siblings, the family size channel would not have been a plausible explanation for

[22] We tried dropping everyone born in the 1920s, then even those born in the 1930s, but the coefficients remained insignificant.

our findings on these outcomes. With respect to fertility, the above-mentioned studies provide no guidance, but as we will see below, we can learn more about that from a different approach.

The other common strategy used to estimate the impact of family size relies on sibling sex composition as an IV. At this point a brief methodological discussion is warranted. Recall from Section 3.2 that a selection bias can arise when we compare children with an older brother to children with an older sister. In addition, the validity of the sex composition instrument is questioned if social mechanisms are present. These considerations suggest to treat results from this strategy with caution. It seems best to focus on reduced form estimates that were obtained on first-borns only and interpret them as the total impact of a younger sibling's gender. Such estimates would provide an interesting comparison for our estimates on the impact of a co-twin's gender.

In this spirit it is worthwhile to discuss a recent paper by Cools and Kaldager Hart (2015) on fertility outcomes. They use sibling sex composition as an IV for family size; as explained above, we focus on reduced form results in this setting. Similarly to the above studies, Cools and Kaldager Hart (2015) look at Scandinavian data, namely Norwegian singletons.[23] They find that first-born men have more children if their second-born sibling is of the same sex, while the gender of the second-born sibling has no impact on the fertility of first-born women. Thus, their results are in the same direction as our findings. Now let us compare the magnitudes: their reduced-form estimate for the fertility of first-born men is a bit below 0.03, whereas it is 0.044 in our twin sample.[24] It seems difficult to argue that family size is affected more by the sex of the twins than by the sex of the first two children. Thus, if our results reflected the impact of family size, the estimates of Cools and Kaldager Hart (2015) should be bigger than ours and not the other way around. In contrast, the smaller estimates of Cools and Kaldager Hart (2015) can be consistent with explanations based on social mechanisms, since those mechanisms are likely to be weaker when they come from younger siblings towards older children.

[23] They briefly look at twin births as well and find that having twin siblings has no significant impact on the fertility of the other siblings. See also Kolk (2015).

[24] We calculated the reduced form estimates from the IV estimates and the first stage results.

Table 3.4. Potential channels on the sample of DZ twins

	Number of siblings	Age at menarche
Panel A: Men		
Same-sex co-twin	-0.020	
	(0.036)	
N	13409	
Panel B: Women		
Same-sex co-twin	-0.015	0.038
	(0.036)	(0.027)
N	14663	14283

Note: The standard errors shown in parentheses are robust and allow clustering by the family of origin. All regressions include a constant and cohort fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1

3.5.2 Testosterone transfer

Another potential explanation that we have to consider is testosterone transfer between twins. This channel is based on the Twin Testosterone Transfer (TTT) hypothesis, which is the assumption that testosterone (T) can transfer between twins in utero. Male fetuses produce much higher levels of T, especially in weeks 10-20 of gestation (Baron-Cohen et al., 2004). Therefore if T could transfer, those with a male co-twin would be exposed to higher prenatal T than those with a female co-twin. This means that the estimates could reflect the impact of prenatal testosterone. This argument is followed in a recent paper by Gielen et al. (2016), who compare the earnings of SS and OS twins to each other. In line with our results, they find that a co-twin brother increases men's earnings.[25] Building on the TTT hypothesis, they interpret the estimates as the impact of prenatal testosterone (T).

The inspiration for the TTT hypothesis comes from animal studies. In case of rodents such as mice and rats, direct measures of prenatal T levels show that T can transfer between littermates (e.g. vom Saal and Bronson, 1980; Even et al., 1992). In line with this, rodents that developed between male fetuses differ from rodents that developed between female fetuses in several aspects (Ryan and Vandenbergh, 2002). Most of the differences relate to genital morphology, physiology and reproductive characteristics, as the primary function of prenatal T is the sexual differentiation of the reproductive system.[26]

[25] Gielen et al. (2016) only look at earnings and they do not differentiate by zygosity. Recall that the inclusion of MZ twins does not change the results on the income of men substantially (see Table 3.5). However, the inclusion of MZ twins does bias the coefficient in case of women such that brothers seem to have a more negative impact. Indeed, this bias is reflected in their estimates on women.

[26] In the presence of high prenatal T, male genitalia develop. In the absence of high prenatal T,

The impetus to the TTT hypothesis was given by Miller (1994), who speculated that T may transfer in case of humans as well and argued that human twins should be studied to uncover potential signs of T transfer.[27] Several papers examined the outcomes of twins in this vein - for reviews, see Cohen-Bendahan et al. (2005a) and Tapp et al. (2011). Both reviews find that the results of this literature lack consistency.

The main idea behind the TTT hypothesis is that extrapolations from animals to humans might be possible. To address this issue, we compare our results on fertility to the results of animal studies on reproduction. The focus of these studies has been on females because they produce much lower levels of prenatal T and hence an extra dose is expected to affect their development more. The findings reveal that females enter puberty later and have lower fertility if they develop between male fetuses instead of female fetuses (Ryan and Vandenbergh, 2002). One might suspect that the results on puberty can be extrapolated, since we found that women with co-twin brothers give birth later than women with co-twin sisters. Fortunately, we can examine this explicitly because we have information on age at menarche (first menstruation). As we can see from Table 3.4, the coefficient is insignificant and its sign is opposite to the prediction of this hypothesis. So women with sisters do not mature earlier biologically, which means that the findings on the timing of puberty do not carry over to our sample. Similarly, there is no effect on the fertility of women.

As far as males are concerned, animal studies produced mixed results. Some researchers found that developing between male fetuses instead of female fetuses leads to enhanced sexual performance (Clark et al., 1992). However, others found significant results in the opposite direction (vom Saal et al., 1983). Given these controversial findings, it is difficult to assess whether our results on men are in line with animal studies.

Our analysis so far found little support for the TTT channel. On the other hand, one can argue that our approach has its shortcomings. Most notably, our evidence is indirect: similarly to the TTT literature, we draw inferences from postnatal outcomes because we do not have information about actual prenatal T levels. While this is clearly a limitation, we note that our conclusion is consistent with medical research that tests the TTT hypothesis directly. In particular, Abeliovich et al. (1984) measured T levels in the amniotic fluid of twin fetuses. They found no evidence of T transfer: T levels were not elevated for twins with a male co-twin.[28]

female genitalia develop, together with the internal reproductive organs (Cohen-Bendahan et al., 2005a).

[27] Although some other researchers expressed similar views already earlier (e.g. Resnick et al., 1993), Miller's study drew the most attention and it is considered to be the seminal paper of the TTT literature.

[28] Their data shows that the mean amniotic T level (pg/ml) of female fetuses with a female (male) co-twin is 113 (105). The mean amniotic T level (pg/ml) of male fetuses with a female (male) co-twin

The importance of direct T measurement was emphasized by Miller (1994) as well, but it seems that he was unaware of the results of Abeliovich et al. (1984). Instead, he based his argument on Meulenberg and Hofman (1991), who found higher maternal T levels in pregnancies with a male fetus. He interpreted this as a sign that T can transfer, at least from the fetus to the mother. However, several other studies failed to find a difference in maternal T levels by the sex of the fetus (Glass and Klein, 1981; Rodeck et al., 1985; van de Beek et al., 2004; Cohen-Bendahan et al., 2005b; Toriola et al., 2011). In fact, studies that measured T in both amniotic fluid and maternal serum found no correlation between the two measures (Rodeck et al., 1985; van de Beek et al., 2004). Hence T transfer between the fetus and the mother seems to be doubtful.

To summarize, we found that the results of the animal literature do not carry over to women and the picture is mixed in case of men. In addition, studies that measure prenatal T levels directly obtain results that are not consistent with the hypothesis of T transfer. Therefore we conclude that there seems to be little support for the TTT channel.

3.5.3 Social mechanisms

As we have seen in the previous two subsections, it is unlikely that the results are led by the family size channel or by testosterone transfer. Let us turn now to a competing explanation. Since siblings are one of the most important family members and peers, they could affect each other via various social mechanisms, such as childhood spillovers, reference point considerations and changes in parental treatment. In this subsection we discuss these social mechanisms in detail and examine whether our results could be explained by them. Unfortunately our data does not allow a straightforward test of the specific social mechanisms, so we will try to evaluate the plausibility of these mechanisms by analyzing whether their predictions are in line with our findings.

First, let us discuss childhood spillovers. Recall that the child development literature finds that those with brothers (sisters) become more masculine (feminine) (Rust et al., 2000). This suggests that gender differences in personality traits and preferences may affect the siblings. For example, men are less risk averse, more competitive, more willing to negotiate, less socially minded, less agreeable and less neurotic than women (see the review of Bertrand, 2010)[29]. If some of these preferences and traits spill over

is 315 (301). The difference between male and female fetuses is significant. The differences by the sex of the co-twin are not significant, neither for males, nor for females (and they are not even in the direction predicted by the TTT hypothesis).

[29] Agreeableness and neuroticism are the two personality traits from the so-called "Big Five" that consistently show gender differences.

to the sibling, it might explain why those with brothers have higher income.[30] Moreover, previous research suggests that the impact will be more pronounced for men. Of these six non-cognitive factors, three have been found to affect the earnings of men and women differently: negotiation, agreeableness and neuroticism. Bowles et al. (2007) find that women who initiate negotiations receive worse evaluations than men who initiate negotiations, at least when the evaluators are men. Similarly, Säve-Söderbergh (2007) finds that employers reward self-promoting less in case of women. Mueller and Plug (2006) find that being less agreeable and less neurotic are both associated with higher earnings in case of men, but not in case of women. Thus, an impact on men's income is more likely and this is indeed what we see.

A second possibility is the reference point argument. Research on adult siblings shows that people view their siblings' outcomes as reference points (Adams, 1999; Keim et al., 2009). Therefore they do not want to lag behind their siblings. This is in line with Kahneman and Tversky (1984): there is loss aversion for outcomes below the reference point. This can apply to labor market outcomes and family life as well. Several recent studies report findings that are consistent with the reference point mechanism. Kuegler (2009) finds that individuals are less satisfied with their life if their sibling earns more than they do.Dahl et al. (2014) find that men are more likely to take paternity leave if their brother was exogenously induced to do so. Joensen and Nielsen (2015) find that high school students are more likely to chose advanced math and science courses if their older sibling unexpectedly could choose math-science at a lower cost.

The gender of the sibling is relevant for two reasons. First, rivalry is stronger among same-sex siblings (Adams, 1999). This is especially true with respect to family formation, since gender roles are quite different in that respect (Keim et al., 2009). This means that we can expect earlier family formation and higher marriage and fertility rates among same-sex twins, and perhaps better labor market outcomes as well. The second reason to expect gender to matter is that gender differences imply different reference points. If the labor market outcomes of men are better, brothers will represent a higher reference point than sisters. Similarly, as women marry and have kids at an earlier age, we can expect sisters to decrease age at first birth.[31]

We consider the predictions of the reference point mechanism clear when same-sex

[30] Such spillovers can occur in several ways. Sex composition affects the type of common play, which can then affect the acquisition of certain traits (Stoneman et al., 1986). Access to toys is also affected, since parents buy sex-typed toys for their children (Fisher-Thompson, 1993; Nelson, 2005). Yet another option is that parental behavior differs by sex composition (e.g. boys might be encouraged more to compete with each other).

[31] The mean age at first marriage is about two and a half years lower for women than for men in Europe (UNECE, 2012). The difference in median age at first marriage is 2.5 years in the US (Copen et al., 2012). Similar gaps can be found in age at first birth, since age at marriage and age at first birth are highly related (see Baizán et al., 2003).

rivalry and gender differences do not go in opposite directions. This means that for earnings, we can only predict the impact on men: those with brothers should earn more. This is consistent with our findings. In terms of marriage and fertility rates, we expect both men and women to perform better if they have a same-sex sibling. However, only women are expected to also give birth earlier. The predictions on timing are in accordance with our results, but the predictions on marriage and fertility are only in line with the findings of men. As far as education is concerned, the prediction on women is not so clear-cut. On the one hand, same-sex rivarly would predict higher education for women with sisters. On the other hand, Butcher and Case (1994) argue that women with sisters obtain lower education because they get a lower educational reference point. Although the gender gap in education closed by now, we can expect this argument to hold for the older cohorts. In line with this, the result on women's education is led by the older cohorts (untabulated results). For boys, the educational prediction is clear: we can expect men with brothers to have higher education. We do not see such a pattern in the data. Perhaps rivalry among brothers is focused on other educational outcomes such as test scores (Nicoletti and Rabe, 2014) or course choices (Joensen and Nielsen, 2015).

Finally, effects can arise indirectly, via differential parental treatment. Previous research shows that parents support the family formation of daughters more than the family formation of sons (Pollet et al., 2009; Danielsbacka et al., 2011). The results of these studies show that grandparents have more contact with the children of their daughters than with the children of their sons. They also provide more informal childcare and are more likely to provide essentials, gifts and extras for the baby. This differential treatment suggests that parents have a preference for grandchildren on their daughter's side. This implies that there is a substitution effect from sisters: if someone has a sister instead of a brother, he/she gets less support (and probably also less pressure) from his parents to form his own family. The prediction from this mechanism is clear: sisters negatively affect the family formation of their siblings. This prediction is consistent with the results on men. On women, we do not see a negative impact. However, recall that the reference point mechanism predicted a positive impact on women. Thus, the null results on women can simply reflect the fact that the two opposing forces cancel each other out.

To sum up, spillovers in childhood, reference point considerations and differential parental investment can all have an impact. Although we could not conduct a straightforward test of these mechanisms, we tried to examine their plausibility by analyzing their predictions. We found that the predictions from childhood spillovers are in line with the results. If we combine the predictions of the reference point argument and the differential parental treatment, the joint predictions are also quite close to what we observe in the data. Therefore we conclude that social mechanisms provide a plausible explanation for the findings.

3.6 Conclusions

Using a sample of dizygotic twins from Sweden, we examined whether the gender of the co-twin has an impact on individuals' education, earnings and family formation. We find that men and women are both affected, but in different ways. Men with brothers have higher earnings, are more likely to form a family and have more children. Women with sisters have lower education and give birth earlier. The effect sizes are small but comparable to the impact of other family factors.[32]

Our analysis showed that social mechanisms are the most likely explanation for the results. In particular, the result that brothers increase men's earnings could be explained by childhood spillovers or reference point considerations in adulthood. The finding that men with brothers are more likely to form a family than men with sisters could be driven by same-sex rivalry or by the fact that parents give more support for the family formation of daughters than for the family formation of sons. The finding that women with sisters have lower education is consistent with Butcher and Case (1994)'s proposed explanation that women used to get lower educational standards if they had a sister instead of a brother. Finally, the result that women with sisters give birth earlier could arise from same-sex rivalry or the fact that sisters provide an earlier reference point for age at first birth.

Can we expect similar effects among singleton siblings? The abovementioned social mechanisms are not twin-specific, so in principle the results could carry over to other samples. However, twins are probably closer to each other, so effects might be smaller between singletons. We can expect the effect to decrease as the difference in age increases. An age difference of four years seems to be a critical threshold; Joensen and Nielsen (2015) find that sibling spillovers in educational choices disappear when the difference between siblings is larger than this. Another consideration is the hierarchy among siblings. Previous research on siblings indicates that there is a leader-follower relationship between older and younger siblings, so older siblings can have a bigger impact on younger siblings than the other way around. Twins are born at the same time, so their status is more equal. This predicts that the impact of a co-twin is smaller than the impact of an older sibling but bigger than the impact of a younger sibling. Taken together, these considerations suggest that our estimates are an upper bound

[32] Holmlund et al. (2011) find that the causal effect of an additional year of parental schooling on children's schooling is around 0.10, which is exactly the magnitude that we find. The causal effect of family size is small and insignificant in most studies (see also Section 3.5.1).

for the impact of a younger sibling's gender. However, the impact of an older sibling's gender might be smaller or bigger than the impact of a co-twin's gender, as long as the age difference is not too big. Our prediction for the impact of younger siblings is in line with the results of Cools and Kaldager Hart (2015). As we discussed in Section 3.5.1, they find that the gender of the second-born sibling affects the fertility of the first-born sibling the same way as the gender of the co-twin affects the other twin. The only difference is that the effect is smaller, just as we expected.

It is worth to discuss external validity in a more general sense as well. Twins are an obvious choice if we want to model two singleton siblings born at the same time. However, one can worry that twins might be too special so that the results do not carry over to the general population. This can be mitigated somewhat by the fact the we look at a time period when IVF technology was unavailable. Nonetheless, the concern remains valid and it also underlines that every empirical strategy has its limitations. Comparing results with studies that use a different strategy can help in this respect. The above-mentioned study by Cools and Kaldager Hart (2015) is a good example since they examined singletons. As their results on fertility are in line with our estimates, it seems that our findings are not specific to twins.

While our study contributes to knowledge about the long-term effects of a sibling's gender, there are several questions that remain open. This is particularly true with respect to the mechanisms behind our results. We found that there are various social mechanisms that are promising candidates to explain the findings. However, our analysis on this topic remained indirect as we do not have direct measures of the underlying variables. This also means that the relative contribution of the different social mechanisms is still an open question. An exciting direction for future research could be to examine settings in which one can conduct more straightforward tests of the various social mechanisms.

Appendix

Table 3.5. Estimates of the effect of a same-sex co-twin on education, income and family formation on all twins, irrespective of zygosity

	Years of education	Ln(perm. income)	Ever married	Any kids	Number of kids	Age at first birth
Panel A: Men						
Same-sex co-twin	0.065	0.014**	0.010**	0.009	0.036*	-0.018
	(0.043)	(0.006)	(0.005)	(0.006)	(0.020)	(0.090)
N	18314	18314	18157	17986	17981	14681
Panel B: Women						
Same-sex co-twin	0.000	0.001	-0.003	-0.008	0.000	-0.045
	(0.038)	(0.008)	(0.005)	(0.005)	(0.018)	(0.077)
N	20371	20358	20271	20117	20115	17464

Note: The standard errors shown in parentheses are robust and allow clustering by the family of origin. All regressions include a constant and cohort fixed effects. Education is measured as years of schooling. Permanent income is average income in years 1971-2007 (if subject is aged 25-64 in the given year). The variable "ever married" takes one if the individual has ever been married or cohabited and zero otherwise. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Chapter 4

The effect of extra school resources on test scores in a Weighted Student Funding system: A nonparametric bounds analysis[1]

4.1 Introduction

Changing school resources is a key policy instrument in education. However, its impact on students' achievement is debated among academics.[2] In this chapter we focus on a policy measure that provides extra funding to schools for students with a disadvantaged family background. More specifically, we look at the Netherlands, where the funding of primary schools is based on a national formula that includes extra weights for disadvantaged students. We examine the causal effect of the extra funding on student's test scores, using a nonparametric partial identification technique.

The Dutch funding scheme is a version of the so-called Weighted Student Funding (WSF) system (Ladd and Fiske, 2011). WSF systems have three main elements: 1) a money-follows-students mechanism whereby the funding of a school depends on the number of students, 2) extra weights are assigned to disadvantaged students and therefore their schools get more resources, and 3) schools have a lot of freedom to use

[1] I am very grateful to Monique de Haan for all her helpful suggestions and for providing me the Stata code that she used to estimate the nonparametric bounds in De Haan (forthcoming). A modified version of that code is used to estimate the nonparametric bounds in this chapter.

[2] Hanushek (2003) and Krueger (2003) are two classic papers in this debate. More recent studies also provide mixed findings. For example, Holmlund et al. (2010), Machin et al. (2010), Hægelanda et al. (2012), Gibbons et al. (2012), Jackson et al. (2016) and De Haan (forthcoming) and find that additional resources have a positive effect, while Bénabou et al. (2009) find no effect. Leuven et al. (2007) and Van der Klaauw (2008) even find negative point estimates, although these estimates are not always significant.

these resources as they find appropriate (Ladd and Fiske, 2011). The proponents of the WSF approach argue that such a funding scheme could foster equal opportunity, as disadvantaged children need help to get the same life chances as those who have a more favorable background (OECD, 2012; Fordham Institute, 2006).

The WSF concept has been appealing to policymakers worldwide. The Dutch were one of the firsts to implement it, as early as 1985.[3] Over time, several US school districts introduced it, including such major cities as New York City, San Fransisco and Boston (Archer, 2005; Furtick and Snell, 2013). In 2006, a prominent conservative think tank argued that the entire public school finance system in the US should be based on WSF (Fordham Institute, 2006). The proposal's signatories included three former U.S. secretaries of education, indicating that WSF is getting popular among policymakers at the highest levels. Similar developments took place in England. In 2011, the Department of Education proposed a comprehensive reform that would bring funding close to WSF (DfE, 2011).[4] They suggested to adopt a new national formula with additional funding for specific causes, such as deprivation. Since complete transition to a new system can take a while, they started the reform by adding a "pupil premium" in the existing setting, which meant extra funding to schools for students with additional educational needs.

Given the growing popularity of the WSF approach, it would important to understand whether and how the additional resources that are provided within this system affect students' performance. However, it is difficult to identify the causal effect since the assignment of weights is not random but is based on the background characteristics of the students. Due to this identification challenge, the causal effect of the extra funding does not become clear from the previous WSF-literature. WSF-studies are typically more informative on topics such as the characteristics of the specific WSF policy, the resulting resource allocation patterns, and the past and present situation of the disadvantaged and non-disadvantaged students. In the Dutch context that we examine, the most important characteristic of the WSF policy is that disadvantaged status is defined by parental education and immigrant status. With respect to the empirical findings, several important observations have been made. Mulder and Van der Werf (1997) analyzed data from the first few years that followed the implementation of the WSF policy. They found a slight decrease over time in the performance gap

[3] The other early implementation was in the Edmonton school district of Canada, in 1980 (Delaney, 1995).

[4] To some extent, disadvantages were already addressed in the existing funding system. However, this was not transparent, as the amounts depended on historical decisions made by previous governments and local authorities. Indeed, the Department of Education notes that "the [current] system is extremely difficult to understand. It is almost impossible to explain why a particular school receives the budget that it does." (DfE, 2011, p. 2) They argue that a reform is necessary because "schools require a system in which funding is transparent; where funding follows the pupil and where pupils with additional needs attract additional funding." (DfE, 2011, p. 2)

between immigrant students in schools with substantial extra resources and the rest of the population. Mulder and Van der Werf (1997) caution against interpreting this as the success of the policy, especially since among these immigrant students there was a parallel increase in the average length of stay in the Netherlands.[5] Studies analyzing later years find that disadvantaged students still lag behind the non-disadvantaged students and that school practices are worse in schools that have many disadvantaged students (see e.g. Roeleveld et al., 2011 and Ladd and Fiske, 2011). Another important empirical finding is that there does not seem to be a negative spill-over effect on test scores from immigrant children to native Dutch children (Ohinata and Van Ours, 2013). In addition, Ladd and Fiske (2011) points out that in the Netherlands, students with low educated parents do better on PISA tests than comparable students in other OECD countries. This could be interpreted as suggestive evidence for the effectiveness of the Dutch WSF policy, but as the authors themselves clarify, such observations are not sufficient to make more explicit, causal statements about the effect of the extra funding.[6]

The main challenge for empirical identification is that treatment assignment depends on parental background, so differences in the performance of the treated and the non-treated group can arise due to several factors. On the one hand, they can reflect the impact of the extra funding. On the other hand, they can also reflect the effect of background characteristics: treated students might perform differently because their family background is different. To isolate the causal impact of the extra funding, we use a nonparametric partial identification method that was developed by Manski (1989), Manski (1997), Manski and Pepper (2000) and Manski and Pepper (2009).[7] We use test scores as outcome variables: scores on a math test, scores on a language test, scores on an information-processing test and a total score that is a combination of the three subscores. Our starting point is that test scores are bounded and therefore we can construct upper and lower bounds around the average treatment effect. These bounds do not require us to make any assumptions; however, they are very wide. To tighten the bounds, we successively layer various nonparametric assumptions. These

[5] It is well established that length of stay affects the performance of immigrants (see e.g. Ohinata and Van Ours, 2012).

[6] In contrast, there are several other Dutch educational policy measures that have been thoroughly evaluated. For example, Leuven et al. (2007) examine the effect of additional, transitory funding that was provided on top of the WSF-funding for schools with a very high share of disadvantaged students (the policy measure that Leuven et al., 2007 examine was not in place in the years that are analyzed in this study). See also De Haan (forthcoming) for the effect of Learning Support, a policy measure that gives additional resources for low-ability students in secondary schools.

[7] The application of this method is becoming increasingly widespread, see for example Gerfin and Schellhorn (2006), Blundell et al. (2007), Kreider and Pepper (2008), Kreider and Hill (2009), Nicoletti et al. (2011), De Haan (2011), Giustinelli (2011), Gundersen et al. (2012), Kreider et al. (2013), Manski and Pepper (2013), Richey (2014), Hof (2014), Mariotti and Meinecke (2015), De Haan and Leuven (2016), Kreider et al. (2016), De Haan (forthcoming), Almada et al. (forthcoming).

assumptions are weaker than the ones utilized in conventional parametric approaches.

The first assumption is the Monotone Treatment Selection (MTS) assumption, which states that the mean potential test scores of the students who are treated are at most as high as the mean potential test scores of the students who are not treated. This is essentially the same assumption as the one that underlies this policy: the students who are treated are disadvantaged. Our second assumption is that we can use the average income in the neighborhood of the schools as a monotone instrumental variable, thereby assuming that there is a non-negative relationship between mean potential test scores and the income level in the neighborhood. The MTS-MIV bounds indicate that the extra funding has a positive impact on math scores, which can exceed 1.9 standard deviations. The bounds around the impact on language scores, information-processing scores and total scores also allow for large positive impacts. However, they also allow for a negative impact, although to a more moderate extent (only for the information-processing test does the bound allow a negative impact that exceeds 0.1 standard deviations). As a negative impact is perceived by most policymakers as highly unlikely, we subsequently show what happens if we layer the Monotone Treatment Response (MTR) assumption on the bounds, which states that the extra funding cannot hurt students. These MTR-MTS-MIV results show a positive impact on all subscores and the total score as well, with all lower bounds being at least 0.1 standard deviations and all upper bounds being at least 1.7 standard deviations. Thus, our results show that when the impact on students is assumed to be non-negative, the average treatment effect is actually significantly positive.

This chapter proceeds as follows: Section 4.2 explains how WSF is implemented in the Netherlands. Section 4.3 describes the data that we use. We explain our empirical approach in detail in Section 4.4 and also present the results in that section. Finally, Section 4.5 summarizes the conclusions.

4.2 The Dutch context

4.2.1 Primary education system

Primary education in the Netherlands starts at age 4 and typically lasts until the age of 12, when students proceed to secondary school.[8] Primary schools can be of three main types: schools with a public profile, schools with a religious background, and schools that apply some specific pedagogical approach (e.g. Montessori, Dalton).[9] Schools

[8] Children can start school when they turn 4 and most of them do indeed attend school from that age. School becomes compulsory on their 5th birthday.

[9] In addition to these regular school types, there are also some special school types for children with special needs (e.g. for children with serious handicaps or disorders). In this study we focus

with a religious background or a specific pedagogical approach are governed by private school boards, whereas schools with a public profile are governed by the municipality. However, all schools are financed primarily by the central government, irrespective of their type. They receive the public funding according to the same national rules, and are subject to the same accountability standards.

Parents can freely choose primary schools for their children. An important condition of the public funding is that schools are not allowed to charge tuition fees. Children get priority in nearby schools. Schools with a religious background or a specific pedagogical approach can refuse students based on differences in beliefs or views, but this is not a common practice and is allowed only if there is a nearby school with a public profile where the student can be placed.

4.2.2 School funding

Schools receive two types of resources from the central government: payment for personnel and money for materials and supplies. In their overview of school funding in the Netherlands, Ladd and Fiske (2011) find that these two funding types together account for about 90 percent of the schools' budget.[10] The central government applies a weighted student approach to both the personnel- and the material-funding.[11] In the time period that we examine (which is the 1998/1999 and 1999/2000 school years), there were four categories of students with extra weights. Of these, the two most important ones were immigrant children with low-educated or low-skilled parents and native Dutch children with low-educated parents. These students had an extra weight of 0.90 and 0.25, respectively. The higher weight of disadvantaged immigrant students reflects that they were considered to be more disadvantaged than native students with low-educated parents (Mulder and Van der Werf, 1997). The other two categories of weighted students were children of shippers and children of caravan dwellers, who had extra weights of 0.40 and 0.70, respectively. These two groups of students received less attention in the policy discussions as they accounted for a very small percentage of

on regular education ("basisonderwijs" in Dutch) and will therefore not discuss the special education sector ("speciaal basisonderwijs" and "speciale scholen" in Dutch).

[10] The remaining 10 percent consists of subsidies from municipalities or related agencies, and of various other revenue (Ladd and Fiske, 2011). Municipalities are not allowed to discriminate against the privately operated schools and hence such subsidies go to all type of schools. The various other revenue sources are parental contributions, private sponsorships and rental income for facilities such as gymnasiums. Parental contributions are voluntary and are used for nonessential extra activities such as school trips or extracurricular school programs. Ladd and Fiske (2011) analyze how these resources relate to student's background. They find that subsidies from the municipality are higher for schools with more disadvantaged students. In contrast, revenue from the other sources is higher for schools with fewer disadvantaged students, so the funding from these two sources more or less counterbalances each other.

[11] We are grateful to Joop Groos from the Dutch Ministry of Education for explaining the characteristics of the system.

Table 4.1. Student categories and their assigned weights

Category	Description	Weight
Category A:	Students whose parents obtained at most a pre-vocational secondary education (VMBO).	1.25
Category B:	Students who live in a foster home and whose father or mother is (or was) a shipper.	1.40
Category C:	Students whose parents are caravan dwellers.	1.70
Category D:	Students with a non-Dutch cultural background who fulfill one of the following criteria: 1. the father obtained at most a pre-vocational secondary education (VMBO). 2. the mother obtained less than a pre-vocational secondary education (VMBO). 3. the parent with the highest earning has a physical or manual job, or neither parents have income from work.	1.90
Category E:	All other students.	1.00

Note: Non-Dutch cultural background means either that the student belongs to the Moluccan ethnic group; or that at least one of the parents is a refugee; or that at least one of the parents is from Turkey, Morocco, Tunisia, Spain, Portugal, Greece, Italy, the former Yugoslavia, the Netherlands Antilles, Aruba, Suriname, Cape Verde or another non-English speaking country outside Europe, with the exception of Indonesia.

the total population. Table 4.1 shows a detailed description of all the categories and weights, based on the official laws and regulations in 1998.[12]

The main idea behind the weights is that they show how much more funding a school gets from the central government for a weighted student (e.g. 90 percent more resources for a 1.90 student). This simple interpretation is how weights are typically discussed in policy debates. However, the funding system is more complicated in practice.[13] Most importantly, the central government assigns the extra funding to schools and not to individual students. Therefore the amount of extra funding is determined by school-level variables. In the followings, we describe how these variables are calculated.

The basic amount of both funding types (personnel and materials) depends on the number of students in the school, irrespective of their weights. More precisely, let N denote the number of students in a school.[14] The regulations describe that to get the

[12] These categories and weights remained essentially the same until 2006, when a comprehensive reform was started. The reform introduced new weights and categories that were based solely on parental education. The new system was phased in gradually over time and included some transitory elements, to avoid sudden large changes in the funding of schools. See Ladd and Fiske (2011) for more details.

[13] We will discuss here only those main elements of the system that are relevant for our analysis. Further details can be found in the original Dutch regulations and in Ladd and Fiske (2011), who give an overview of this topic.

[14] The number of students and their weights are recorded every year on the 1st of October. To determine funding in a school year, the records of the previous year are used.

number of students that counts for the basic funding, the actual number of students has to be multiplied by 1.03, and then the resulting number should be rounded down to the nearest integer. That is, the number of students that count for the basic funding can be calculated as

$$N_1 = INT\{1.03 * N\} \tag{4.1}$$

The regulations contain correspondence tables that show how much funding is granted for each possible value of N_1. Funding for personnel is expressed in terms of full-time equivalent units, and funding for materials is expressed in terms of number of classes. Other than the unit names, the tables for the two type of funding are essentially the same, so for each N_1, the number of personnel units equals the number of classes. For example, if $N_1 = 60$, the government pays the salary of 3.0 full-time equivalent personnel units, and gives money to cover the materials of 3.0 classes. The correspondence tables show that the basic funding is a monotonically increasing, non-linear function of N_1.[15]

On top of the basic funding, schools with many disadvantaged students get extra funding of both types to combat educational disadvantages. The entitlement for the extra funding is calculated in several steps. First, the weights of all students in the school are added together. Then this sum is reduced by 9% of the number of students. This reduction was built in the policy because of budgetary considerations and was later justified on the ground that schools should be able to cope with the challenge of educating a few disadvantaged students without extra resources (see Ladd and Fiske, 2011). After the resulting number is rounded down to the nearest integer, it is called the weighted number of students in the school. If this weighted number of students is smaller than the actual number of students, the actual number of students is used in the next steps; otherwise the weighted number of students is used. The appropriate number is multiplied by 1.03 and then the resulting number is rounded down to the nearest integer. Let us denote this new variable by N_2. If we denote the weight of student i by w_i, we can summarize the entire calculation process in the following formula:

$$N_2 = INT\left\{1.03 * max\left[INT\left(\sum w_i - 0.09N\right); N\right]\right\} \tag{4.2}$$

Once the N_2 of a school is known, it can be translated to funding units, using the same correspondence tables as the ones used for the basic funding. If this results in the same number of funding units as the basic funding, the school does not get any additional resources. However, if it results in a higher number of funding units, the

[15] In particular, funding increases by 0.2-0.3 units as N_1 increases by 6-11 students. Dobbelsteen et al. (2002) exploits this stepwise function as an instrument for class size.

school gets extra funding. For personnel, the amount of the extra funding equals the difference between the N_2-based funding units and the N_1-based funding units. For materials, the amount of the extra funding equals half of the difference between the N_2-based funding units and the N_1-based funding units. While the magnitude of the additional funding differs between the two types, the concordance of the tables implies that the schools that get extra personnel units are exactly the same as the schools that get extra money for materials.

The policy aim behind the extra funding was to improve the educational opportunities of disadvantaged children, with an emphasis on raising achievement in math and Dutch language (Mulder and Van der Werf, 1997). While this was widely discussed in policy debates, it appears only in a limited form in the regulations. In particular, the regulations state that the aim of the extra funding is to combat educational disadvantages, but they do not give any guidelines or prescriptions on how the extra funding should be used. Importantly, the regulations do not say anything about targeting, so schools are not asked to spend the resources specifically on the weighted students. This means that all students whose school receives extra funding can benefit from it. Schools are also not required to spend the funding in a specific way. For example, the extra personnel units can be used for hiring extra teachers such as regular classroom teachers, remedial teachers or academic coaches. Another possibility is to hire various support staff, such as assistant teachers. From the study of Ladd and Fiske (2011) it seems that the extra funding is indeed utilized in both ways: schools that receive extra funding have more teachers per student and also more support staff.

4.3 Data

We combine various datasets for the analysis. To measure outcomes, we use the standardized scores of the so-called “Cito” test from the 1998/1999 and 1999/2000 school years.[16] Cito is a nationwide test that is taken at the end of primary education. Together with the advice of the primary school teachers, the scores on this test determine the secondary school track in which the student can subsequently enroll. The test is taken by most students, although it is not compulsory.[17] The total Cito score is based

[16] We first standardize the test scores by year and then pool the data.

[17] Some educators argue that tests are too limited in their scope (e.g. they do not capture socioeconomic development) and therefore some schools prefer to use their own methods to evaluate students. Such arguments are especially popular in schools with a specific pedagogical approach (e.g. Montessori and Vrije schools), which serve more advantaged students. In turn, schools that do not participate in the Cito testing have a lower share of disadvantaged students. The government tried to encourage universal participation at the end of the nineteen-nineties, and the share of Cito-taking schools did indeed increase somewhat, from 78% in 1998/1999 to 80% in 1999/2000. In the Appendix we present results separately for the two years, in Table 4.4 and 4.5. The results show that the bounds around the average treatment effect are higher in 1999/2000.

on performance on three tests: a math test, a Dutch language test and an information processing test.[18] The math and the language test is based on the content of the corresponding subjects in primary school. The information processing test is not subject-specific; instead, it measures how well the student can understand information from various sources, such as texts, tables, graphs, diagrams and maps. Since policymakers were particularly concerned about performance in math and Dutch language, we will look not only at the total Cito score but also at performance on the three tests separately.

We merge the Cito dataset to school level administrative data that contains the number of students per weight category. The Cito dataset does not contain information about student weights, so we do not know the weights at the individual level. Nonetheless, this information is not necessary to find out treatment status, since that depends on school level variables. In particular, recall from Section 4.2 that the extra funding is determined at the school level, and all students whose school receives extra funding can benefit from it. Therefore all students who attend schools with extra funding are assigned to the treated group, and all students who attend schools with no extra funding are assigned to the non-treated group.

Finally, we use a third dataset which contains geocoded information on average neighborhood disposable income per person in 1998. As we will explain in Section 4.4.2, we will use this data for our Monotone Instrumental Variable (MIV) assumption. Disposable income is calculated by Statistics Netherlands such that taxes and premiums are subtracted from people's total income. Neighborhoods are geographical areas defined by local authorities that residents typically view as meaningful entities. The Netherlands was divided into 10737 neighborhoods on the reference date of our neighborhood dataset. Using a geographic information system (namely, QGIS) we locate to which neighborhood each school belongs to. We will use average income in the neighborhood of the school as a Monotone Instrumental Variable (MIV) in Section 4.4.2.[19]

We present descriptive statistics by treatment status in Table 4.2. As we can see from the table, the standardized test scores of treated students are significantly worse

[18] In addition to these main parts, there is an additional, optional test called "world orientation", which is a test in geography, history and natural sciences. This part does not contribute to the total Cito score, so we do not use it in the analysis.

[19] As neighborhoods are defined by many different local authorities in the country, they can be of various sizes. When a neighborhood is too small to fill its school with students, we calculate the average income in the neighborhood of the school as the weighted average of the income in the own neighborhood and of the income in the next closest available neighborhood. When calculating this average, we weight with the expected number of students from the given neighborhood (e.g. if the own neighborhood of the school is so small that it can supply only 20 percent of the students and the remaining 80 percent is probably coming from the next closest neighborhood, we weight the contribution of the own neighborhood by 0.2 and the contribution of the other neighborhood by 0.8).

Table 4.2. Descriptive statistics

	Treated		Non-treated		Difference	
	Mean	S.d.	Mean	S.d.	Mean[a]	S.e.[b]
Test scores						
Total Cito score	-0.199	1.044	0.161	0.933	-0.360***	0.012
Math score	-0.154	1.035	0.124	0.953	-0.278***	0.011
Language score	-0.182	1.033	0.147	0.947	-0.329***	0.011
Info-processing score	-0.210	1.058	0.169	0.916	-0.378***	0.011
Characteristics of student's school						
Share of 1.25 students in school	0.226	0.137	0.125	0.110	0.102***	0.004
Share of 1.40 students in school	0.002	0.016	0.000	0.004	0.001***	0.000
Share of 1.70 students in school	0.004	0.015	0.001	0.005	0.003***	0.000
Share of 1.90 students in school	0.272	0.268	0.019	0.026	0.252***	0.006
Extra funding (as % of basic funding)	21.2	22.2	0	0	21.2***	0.5
Average neighborhood income (€)	9523	1161	10160	1424	-637***	39
Number of observations	121602		149153		270755	

Notes: [a] Significance levels: *** p<0.01, ** p<0.05, * p<0.1 (based on t-statistics). [b] Clustered robust standard errors.

than the standardized test scores of non-treated students. The difference is largest for the information-processing score and smallest for the math score. Overall, treated students score 0.36 standard deviation lower on the Cito than non-treated students. From the bottom part of the table we can also see that in the treatment group the share of disadvantaged students in the school is significantly higher, from all four types. The table also shows that on average the extra funding is approximately 21 percent of the basic funding.

4.4 Empirical approach and findings

Our aim is to identify the causal effect of the extra funding on the test scores of the students. To formalize the empirical problem, we use a potential outcome framework. The basis of this framework is that we define alternative states to which all students could be potentially exposed. Since we are interested in the effect of the extra funding, we define the treatment variable t such that $t = 1$ indicates the state in which a student's school receives extra funding and $t = 0$ indicates the state in which the student's school does not receive extra funding. Note that when we switch t from 1 to 0, we still talk about the same student in the same school, so the only difference is the receipt of the extra funding. With this notation we can make "what if" type of statements: we use $t = 1$ to talk about what would happen if the student's school received extra funding and we use $t = 0$ to talk about what would happen if the student's school did not receive extra funding.

Using this notation we can write the average treatment effect as

$$ATE = E[y(t = 1)] - E[y(t = 0)] \tag{4.3}$$

where y denotes the outcome variable, which is the relevant standardized test score in our case (can be the total Cito score, or the subscore on math, language or information-processing).[20] In words, the average treatment effect is the difference between the mean potential outcome that would occur if everyone was treated and the mean potential outcome that would occur if no one was treated. Since switching t from 1 to 0 keeps everything constant except the extra funding, this difference can be interpreted as the causal effect of the extra funding.

The empirical challenge that we face is that we cannot observe the terms in equation (4.3). To facilitate the discussion on observable and unobservable terms, let us introduce the dummy variable d which indicates realized treatment status, that is,

[20] Since test scores are student level outcomes, we conduct the analysis at the student level. Nonetheless, the results would be qualitatively the same if we collapsed the data and run the analysis at the school level.

treatment status in the situation that we observe. Thus, $d = 1$ for students whose school currently receives extra funding and $d = 0$ for students whose school currently does not receive extra funding. We will occasionally use the shorter notation d^1 for students with $d = 1$ and d^0 for students with $d = 0$. Note that d^1 and d^0 denote two different group of students: the former denotes the students in the treated group and the latter denotes the students in the non-treated group. As we can see from Table 4.2, the two groups of students are different not only in the receipt of treatment, but also in that the share of disadvantaged students is significantly higher in the treated group. Thus, d^1 students are on average more disadvantaged and have more disadvantaged peers than d^0 students.

Using this notation we can clarify when potential outcomes can be observed: if and only if $t = d$. That is, we can observe $y(t = 1)$ only for those students who have $d = 1$ but not for those students who have $d = 0$. Similarly, we can observe $y(t = 0)$ only for those students who have $d = 0$ but not for those who have $d = 1$. Thus, instead of observing $E[y(t = 1)]$ and $E[y(t = 0)]$, we can only observe $E[y(t = 1)|d = 1]$ and $E[y(t = 0)|d = 0]$. The relationship between the ATE and these new terms can be made clear if we make use of the law of iterated expectations and rewrite equation (4.3) as

$$ATE = E[y(t = 1)|d = 1] \cdot P(d = 1) + E[y(t = 1)|d = 0] \cdot P(d = 0) -$$

$$[E[y(t = 0)|d = 1] \cdot P(d = 1) + E[y(t = 0)|d = 0] \cdot P(d = 0)] \tag{4.4}$$

Besides the aforementioned terms $E[y(t = 1)|d = 1]$ and $E[y(t = 0)|d = 0]$, we also know the terms $P(d = 1)$ and $P(d = 0)$ since we we know the realized treatment status of the students. Thus, the only two terms that are unknown from equation (4.4) are $E[y(t = 1)|d = 0]$ and $E[y(t = 0)|d = 1]$. Therefore if we want to learn more about the ATE, we need to learn more about these terms. To this end, we will construct upper and lower bounds around them, and subsequently around the ATE.

Our first step is to clarify what we know without making any assumptions. Our outcome variables are test scores, which are bounded by a maximum and a minimum value. This implies that the unobserved terms are also bounded. That is, we know that

$$y_{min} \leq E[y(t = 1)|d = 0] \leq y_{max} \tag{4.5}$$

$$y_{min} \leq E[y(t = 0)|d = 1] \leq y_{max} \tag{4.6}$$

As Manski (1989) pointed out, we can use such information to construct upper and lower bounds around the ATE. We will call these "NO Assumption" (NOA) bounds. From equations (4.4), (4.5) and (4.6) we can see that the highest possible value of ATE is when $E[y(t=1)|d=0]$ is at its maximum and $E[y(t=0)|d=1]$ is at its minimum. Similarly, we can see that the lowest possible value of ATE is when $E[y(t=1)|d=0]$ is at its minimum and $E[y(t=0)|d=1]$ is at its maximum. Thus, the NOA bounds around ATE can be expressed as

$$E[y(t=1)|d=1]\cdot P(d=1)+y_{min}\cdot P(d=0)-[y_{max}\cdot P(d=1)+E[y(t=0)|d=0]\cdot P(d=0)]$$

$$\leq ATE \leq \tag{4.7}$$

$$E[y(t=1)|d=1]\cdot P(d=1)+y_{max}\cdot P(d=0)-[y_{min}\cdot P(d=1)+E[y(t=0)|d=0]\cdot P(d=0)]$$

We calculate the sample analogs of all the terms from the data and present the results in column (1) of Table 4.3. The other columns of the table are based on different assumptions and will be discussed later, so for now we will focus on column (1) only. For each outcome variable, the first line shows the NOA bounds and the second line shows the confidence intervals around the bounds (the lower confidence interval for the lower bound and the upper confidence interval for the upper bound). The confidence intervals are calculated by a bootstrap procedure with 1000 replications, using the method that was developed for this bounds approach by Imbens and Manski (2004). We take into account clustering at the school level.

As we can see from column (1), the NOA bounds allow for a large negative impact as well as a large positive impact: the size of the effect could be well above 2 standard deviations, in both directions. Although these bounds are too wide to be informative in themselves, they are important because they provide the basis for our subsequent analysis. As Manski and Pepper (2000) and Manski and Pepper (2009) show, we can tighten the bounds by making various nonparametric assumptions. In the following subsections we will apply three types of assumptions: the Monotone Treatment Selection (MTS) assumption, the Monotone Instrumental Variable (MIV) assumption and the Monotone Treatment Response (MTR) assumption.

Table 4.3. Results based on NO Assumptions, on the MTS assumption, on the MTS-MIV assumptions and on the MTR-MTS-MIV assumptions

	(1)	(2)	(3)	(4)
	NOA	MTS	MTS-MIV	MTR-MTS-MIV
Total Cito score	[-2.789 ; 2.243]	[-0.360 ; 2.243]	[-0.054 ; 1.782]	[0.137 ; 1.782]
	[-2.814 ; 2.267]	[-0.378 ; 2.267]	[-0.114 ; 1.949]	[0.089 ; 1.949]
Math score	[-2.908 ; 2.416]	[-0.278 ; 2.416]	[0.035 ; 1.934]	[0.134 ; 1.934]
	[-2.935 ; 2.442]	[-0.295 ; 2.442]	[-0.026 ; 2.112]	[0.083 ; 2.112]
Language score	[-3.284 ; 2.716]	[-0.329 ; 2.716]	[-0.061 ; 2.178]	[0.126 ; 2.178]
	[-3.314 ; 2.746]	[-0.347 ; 2.746]	[-0.125 ; 2.374]	[0.076 ; 2.374]
Information-processing score	[-3.530 ; 2.875]	[-0.378 ; 2.875]	[-0.108 ; 2.303]	[0.101 ; 2.303]
	[-3.563 ; 2.908]	[-0.397 ; 2.908]	[-0.159 ; 2.514]	[0.059 ; 2.514]
Number of observations	270755	270755	270755	270755
Number of clusters	5684	5684	5684	5684

Note: For each outcome variable and assumption combination, the first line shows the estimated bounds and the second line shows the 95% confidence intervals around the bounds. In column (3) and (4) the estimated bounds and confidence intervals are bias-corrected, following the method of Kreider and Pepper (2007) that was developed for MIV-based bounds. All confidence intervals are based on 1000 bootstrap replications, using the method that was developed by Imbens and Manski (2004). We take clustering at the school level into account.

4.4.1 Monotone Treatment Selection (MTS)

Our first assumption is the so-called Monotone Treatment Selection (MTS) assumption. This means that we assume that the mean potential outcomes of d^1 students are not higher than the mean potential outcomes of d^0 students. This assumption is based on the argument that disadvantages have a non-positive effect on performance and that the share of disadvantaged students is higher among d^1 students than among d^0 students. Essentially, we assume that the policy is targeted properly, which seems reasonable especially since the weight system is based on previous research that documented lower educational performance for disadvantaged students (Ladd and Fiske, 2011). Formally, our MTS assumption is expressed by the following two inequalities:

$$E[y(t=1)|d=1] \leq E[y(t=1)|d=0] \tag{4.8}$$

$$E[y(t=0)|d=1] \leq E[y(t=0)|d=0] \tag{4.9}$$

Equation (4.8) states that if all the students were treated, the average test score of the d^1 students would not be higher than the average test score of the d^0 students. Equation (4.9) makes a similar statement for the case when nobody receives treatment. That is, it states that if nobody received the treatment, the average test score of the d^1 students would not be higher than the average test score of the d^0 students.

Note that the MTS assumption is about selection into the treatment and control group, not about treatment response. In particular, the MTS assumption does not claim that the treatment has a positive impact. The MTS assumption allows for the possibility that the treatment has a negative effect on some or even all of the students. What the MTS assumption states is that if everyone received the same treatment, d^1 students would on average not perform better than d^0 students.

The MTS assumption can tighten the bounds around ATE because it imposes new restrictions on the unobserved terms $E[y(t=1)|d=0]$ and $E[y(t=0)|d=1]$. More specifically, the new minimum of $E[y(t=1)|d=0]$ is $E[y(t=1)|d=1]$ and the new maximum of $E[y(t=0)|d=1]$ is $E[y(t=0)|d=0]$. Thus, instead of equation (4.7), we have now new bounds around the ATE:

$$E[y(t=1)|d=1]\cdot P(d=1)+E[y(t=1)|d=1]\cdot P(d=0)-$$

$$[E[y(t=0)|d=0]\cdot P(d=1)+E[y(t=0)|d=0]\cdot P(d=0)]\leq ATE\leq \tag{4.10}$$

$$E[y(t=1)|d=1]\cdot P(d=1)+y_{max}\cdot P(d=0)-[y_{min}\cdot P(d=1)+E[y(t=0)|d=0]\cdot P(d=0)]$$

Note that both of these changes affected the lower bound of the ATE. We can simplify equation (4.10) to

$$E[y(t=1)|d=1]-E[y(t=0)|d=0]$$

$$\leq ATE\leq \tag{4.11}$$

$$E[y(t=1)|d=1]\cdot P(d=1)+y_{max}\cdot P(d=0)-[y_{min}\cdot P(d=1)+E[y(t=0)|d=0]\cdot P(d=0)]$$

Thus, the new lower bound of the ATE is the difference between the test scores of d^1 students and d^0 students. The upper bound is not affected by the MTS assumption. Column (2) of Table 4.3 shows the bounds and confidence intervals that result from applying the method to our data. As we can see, the upper bounds are indeed the same as the corresponding NOA upper bounds, whereas the lower bounds are indeed the same as the difference that we could observe between the treated and non-treated groups in Table 4.2.

4.4.2 Monotone Instrumental Variable (MIV)

The above bounds can be tightened further by using the Monotone Instrumental Variable (MIV) assumption (see Manski and Pepper, 2000). We can use a variable z as an MIV if mean potential outcomes are non-decreasing in z. Formally, if for each $v_1\leq v\leq v_2$ it holds that

$$E[y(t=1)|z=v_1]\leq E[y(t=1)|z=v]\leq E[y(t=1)|z=v_2] \tag{4.12}$$

and

$$E[y(t=0)|z=v_1]\leq E[y(t=0)|z=v]\leq E[y(t=0)|z=v_2] \tag{4.13}$$

The MIV that we use in this study is the average income in the neighborhood of the school.[21] This means that we assume that mean potential test scores are non-decreasing in average neighborhood income. We argue that this is a reasonable assumption because students typically live close to their primary school, which means that those who attend a school in a high-income neighborhood are probably from a high-income neighborhood and family themselves. It is well established that both of these factors are positively associated with educational outcomes (see Haveman and Wolfe, 1995, Klebanov et al., 1998 and Blanden and Gregg, 2004 for reviews and further analysis). Note that the MIV assumption is weaker than the commonly applied IV assumptions, in that IV assumptions impose mean independence, whereas the MIV assumption allows for a weakly monotone positive relation between average neighborhood income and mean potential test scores.

To calculate the MIV bounds around the ATE, we first need to calculate the MIV bounds around the mean potential outcomes. This can be done in four steps. In Step 1, we create subsamples for each value of z. In Step 2, we calculate the lowest and the highest possible value of the mean potential outcomes within each subsample. We combine the MTS and the MIV assumptions such that at this step we apply the MTS assumptions of equation (4.8) and equation (4.9) within each subsample. In Step 3, we apply equations (4.12) and (4.13) to tighten these lower and upper bounds. To see how this works, note that equation (4.12) implies that the upper bound of $E[y(t=1)|z=v]$ cannot be higher than the upper bound of $E[y(t=1)|z=v_2]$. In fact, it cannot be higher than the upper bound in any other subsample where the value of z is higher than or equal to v. Similarly, the lower bound of $E[y(t=1)|z=v]$ cannot be lower than the lower bound of $E[y(t=1)|z=v_1]$ or more precisely, than the lower bound in any other subsample where the value of z is lower than or equal to v. Thus, we can tighten the bounds of the subsamples by replacing the upper (lower) bound of $E[y(t=1)|z=v]$ with the lowest (highest) upper bound that occurs in any subsample in which the value of z is higher (lower) than or equal to v. We can utilize equation (4.13) in a similar fashion to tighten the bounds around $E[y(t=0)|z=v]$ in each subsample. Once we do this process over all values of z, we get the new, MIV bounds in each subsample.[22] In Step 4, we take the weighted average of the sub-sample MIV bounds to get the aggregate MIV bounds of the mean potential outcomes. To get the contribution of a subsample, we take the MIV bound in that subsample and multiply it with the share of students in that subsample. We add the contribution of each subsample together to get the aggregate MIV bounds. Note that we need to calculate

[21] A similar MIV has been used by De Haan (forthcoming).

[22] These upper and lower bounds should not cross, as that would violate our assumptions (see also Blundell et al., 2007). Therefore we checked whether our lower bounds are lower than (or at most equal to) the resulting upper bounds in each subsample, and we found that this is indeed the case.

the aggregate MIV bound for each type of bound, that is, we do the aggregation four times: we calculate aggregate lower bounds around $E[y(t = 1)]$, aggregate lower bounds around $E[y(t = 0)]$, aggregate upper bounds around $E[y(t = 1)]$ and aggregate upper bounds around $E[y(t = 0)]$. All the resulting aggregate MIV bounds are weighted averages of the corresponding subsample MIV bounds (see also Manski and Pepper, 2000).

After we have the aggregate upper and lower MIV bounds around $E[y(t = 1)]$ and $E[y(t = 0)]$, it follows from equation (4.3) that we can calculate the upper and lower MIV bounds around the ATE as

$$\sum_{v \epsilon V} P(z = v) \cdot [max_{v1 \leq v} LB_{E[y(t=1)|z=v1]}] - \sum_{v \epsilon V} P(z = v) \cdot [min_{v2 \geq v} UB_{E[y(t=0)|z=v2]}]$$

$$\leq ATE \leq \tag{4.14}$$

$$\sum_{v \epsilon V} P(z = v) \cdot [min_{v2 \geq v} UB_{E[y(t=1)|z=v2]} - \sum_{v \epsilon V} P(z = v) \cdot [max_{v1 \leq v} LB_{E[y(t=0)|z=v1]}]$$

where LB means lower bound and UB means upper bound. Since we used MTS bounds in step 2, we will refer to the resulting bounds as MTS-MIV bounds.

As Manski and Pepper (2000) and Manski and Pepper (2009) point out, estimates of MIV-based bounds have finite-sample bias so that the estimates tend to be narrower than the true bounds. They show with Monte Carlo simulations that the bias increases as the size of the sub-samples decreases. We take two measures to address this problem. First, we use a minimum size rule to divide the sample into sub-samples, which means that each subsample has to contain at least 2000 observations.[23] Second, we apply the bootstrap bias-correction method that was developed by Kreider and Pepper (2007). In column (3) of Table 4.3 we report the bias-corrected estimates of the MTS-MIV bounds, and the confidence intervals are also based on this bias-correction.

As we can see from column (3) of the table, the combined MTS-MIV bounds are substantially narrower than the previous bounds. In case of the mathematics score, the lower bound is now above 0, so it seems that the effect on the math score is positive (although note that the confidence interval includes values slightly below zero). Notice that the upper bound is still quite high, allowing the effect to be more than 1.9 standard deviations. The lower bounds on the other two subscores and on the overall Cito score

[23] In the Appendix we show results for the 1998/1999 and 1999/2000 school years separately. In those estimations we use a minimum size rule of 1000 observations in each sub-sample.

are still below zero, which means that they allow for a negative impact. For the language score and for the total Cito score the estimated lower bound is around -0.05/-0.06, while for the information-processing test it is a bit lower, around -0.11. Note that the upper bounds are very high for all test scores, and they are actually the highest for information-processing. Thus, the bounds around the information-processing score are not lower than the bounds around the other components, but wider. It seems that our method is less successful in determining the impact on the information-processing score, but this does not necessarily mean that the impact on this component is lower than the impact on other components.

4.4.3 Monotone Treatment Response (MTR)

Another well-established assumption in the literature on bounds is the so-called Monotone Treatment Response (MTR) assumption. It was introduced by Manski (1997) and it assumes that response functions are weakly increasing in the treatment. Formally, this can be written as

$$y(t = 0) \leq y(t = 1) \tag{4.15}$$

We note that this is a rather strong assumption in our context, as it states that the extra funding does not hurt the students' test scores. This means that MTR-based bounds rule out a negative impact by assumption, but not a zero impact. This receives some support in case of math scores from our MTS-MIV bounds, but the situation is more ambiguous in case of the total Cito score and the other subscores. Caution is especially warranted in case of the information-processing test, where the MTS-MIV bounds were wide. Nonetheless, it is interesting to see what happens if one is willing to assume that the extra funding is not harmful. This is motivated by the observation that the discussion on the effect of the Dutch policy measure is centered around the question whether and to what extent it helps, and the possibility that it could have a negative effect on students is rarely discussed (see e.g. Roeleveld et al., 2011). In addition, the fact that the WSF policy has been in place for a long time and survived several changes in the government suggests that many policymakers make the implicit assumption that the extra funding is at least not harmful.

For our application, we need to consider what such an assumption would mean in terms of mean potential outcomes. This is expressed in the following equations:

$$E[y(t = 0)|d = 1] \leq E[y(t = 1)|d = 1] \tag{4.16}$$

$$E[y(t = 0)|d = 0] \leq E[y(t = 1)|d = 0] \tag{4.17}$$

We will combine these assumptions with the MTS and MIV assumptions. Recall that the MIV bounds around the mean potential outcomes were calculated in several steps, and we combined the MTS and the MIV assumptions such that we applied the MTS assumptions of equation (4.8) and (4.9) in Step 2 of this process. We include now the MTR assumption such that we also apply equations (4.16) and (4.17) in that step, which means that we apply these assumptions within each MIV-subsample. The combined MTR-MTS-MIV bounds can now be calculated simply by proceeding with the next steps in the MIV-bound calculation process.[24] After we calculated the aggregate MTR-MTS-MIV bounds around ATE with this procedure, we apply the MTR assumption again if necessary, this time to the final bounds. That is, if the resulting aggregate bounds are still negative, we set the lower bound to 0, since negative values are excluded by the MTR assumption. The final MTR-MTS-MIV bounds around ATE are expressed in equation (4.18) below, which differs from the MTS-MIV bound of equation (4.14) in two respect. First, this time the LB and UB terms are calculated not only with the use of the MTS but also with the use of the MTR assumptions. Second, the lower bound cannot be negative.

$$max\left\{\sum_{v\epsilon V} P(z=v)\cdot[max_{v1\leq v}LB_{E[y(t=1)|z=v1]}] - \sum_{v\epsilon V} P(z=v)\cdot[min_{v2\geq v}UB_{E[y(t=0)|z=v2]};0\right\}$$

$$\leq ATE \leq \tag{4.18}$$

$$\sum_{v\epsilon V} P(z=v)\cdot[min_{v2\geq v}UB_{E[y(t=1)|z=v2]} - \sum_{v\epsilon V} P(z=v)\cdot[max_{v1\leq v}LB_{E[y(t=0)|z=v1]}]$$

As mentioned in the previous subsection, estimates of MIV-based bounds have finite-sample bias. Therefore we apply the bias-correction method of Kreider and Pepper (2007) to the estimates of the MTR-MTS-MIV bounds as well. The results are presented in column (4) of Table 4.3. As we can see, all the MTR-MTS-MIV bounds are positive. All lower bounds are at least 0.1 standard deviations and all upper bounds are at least 1.7 standard deviations. Thus, if the impact on students is assumed to be non-negative, the average treatment effect is actually significantly positive.

[24] As in the case of the MTS-MIV bounds, during these calculations we check whether the upper and lower bounds cross in any of the subsamples, and we find that this is not the case.

4.5 Discussion and conclusions

In this chapter we used a nonparametric method to examine how test scores are affected by additional funding that is provided in the Dutch WSF system. We obtained the upper and lower bounds around the average treatment effect under various assumptions. First we applied the Monotone Treatment Selection (MTS) assumption, which states that the mean potential test scores of the students who are treated are at most as high as the mean potential test scores of the students who are not treated. We argued that this assumption is reasonable because disadvantages presumably have a non-positive effect on performance. Our second assumption was that we can use the average income in the neighborhood of the schools as a Monotone Instrumental Variable (MIV). This means that we assumed that there is a non-negative relationship between mean potential test scores and the income level in the neighborhood. We find this assumption reasonable because those who attend a school in high-income neighborhood are presumably from a high-income neighborhood and family themselves, and both of these factors are positively associated with schooling outcomes (Haveman and Wolfe, 1995, Klebanov et al., 1998, Blanden and Gregg, 2004).

Under the combined MTS-MIV assumption, the estimated lower bounds indicated that the extra funding has a positive impact on the math score (although the confidence interval still included values slightly below zero). The upper bound allowed the effect on the math score to be over 1.9 standard deviations. The bounds around the impact on language scores, information-processing scores and total scores also allowed for similarly large positive impacts. However, they also allowed for a negative impact, although to a smaller extent. Only for the information-processing test did the bound allow a negative impact that exceeds 0.1 standard deviations. We also noted that the upper bound is highest for the information-processing test, which means that the bounds around this subscore are not lower than the bounds around the other components, but wider. Thus, our method seems to be less successful in determining the impact on the information-processing score, but this does not necessarily mean that the impact on this component is lower than the impact on other components.

We subsequently showed what happens if we layer the Monotone Treatment Response (MTR) assumption on the bounds. This assumption states that the extra funding cannot hurt students. This means that MTR-based bounds rule out a negative impact by assumption, but not a zero impact. While we find this a rather strong assumption, we showed what can be learned if one imposes it, because many policymakers seem to make such an assumption implicitly. We found that under the MTR-MTS-MIV assumption all lower bounds are at least 0.1 standard deviations and all upper bounds are at least 1.7 standard deviations. Thus, our results show that when the impact on students is assumed to be non-negative, the average treatment effect is actually significantly positive. Recall that on average the extra funding amounts to about 21 percent of the basic funding, so effects above 0.1 standard deviation would not be considered trivially small in the literature on the effect of extra resources.[25] However, the assumptions on which these bounds are based would also not be considered trivial in this literature. It is not unprecedented to find negative point estimates for the effect of additional resources (see e.g. Van der Klaauw, 2008 and Leuven et al., 2007). Therefore we would caution against drawing strong conclusions from the MTR-based bounds.

What can we learn from the results of the more credible MTS-MIV bounds? In case of the math score, the estimated bounds are above zero, so we get an indication that the effect was positive. However, in case of the total score, the language score and the information-processing score, the lower bounds are below zero and the upper bounds are far above it, so we do not know the sign of the effect. These bounds are clearly less informative than one would have hoped for. Nonetheless, given that the causal effect was largely unknown so far, these MTS-MIV bounds still contribute to our knowledge. Recall that without making any assumptions, the only thing that we could say about the average treatment effect is that it is somewhere between minus 2.8 standard deviations and plus 2.2 standard deviations in case of the total score, and it is on an even wider range in case of the subscores. The MTS-MIV bounds reduced the range substantially for all the scores, so in that sense we did learn a lot from them about the effect. The most important lesson was that we can exclude a large negative impact on the total score and on the other scores as well.

[25] For example, Holmlund et al. (2010) find that test scores increase by 0.05 standard deviation due to a 33% increase in resources, and view this favorably.

Appendix

Table 4.4. Results based on NO Assumptions, on the MTS assumption, on the MTS-MIV assumptions and on the MTR-MTS-MIV assumptions, in the schoolyear 1998/1999

	(1)	(2)	(3)	(4)
	NOA	MTS	MTS-MIV	MTR-MTS-MIV
Total Cito score	[-2.762 ; 2.270]	[-0.364 ; 2.270]	[-0.098 ; 1.699]	[0.120 ; 1.699]
	[-2.789 ; 2.297]	[-0.385 ; 2.297]	[-0.164 ; 1.881]	[0.074 ; 1.881]
Math score	[-2.880 ; 2.444]	[-0.290 ; 2.444]	[0.001 ; 1.845]	[0.140 ; 1.845]
	[-2.910 ; 2.474]	[-0.310 ; 2.474]	[-0.064 ; 2.033]	[0.090 ; 2.033]
Language score	[-3.060 ; 2.571]	[-0.336 ; 2.571]	[-0.110 ; 1.914]	[0.104 ; 1.914]
	[-3.092 ; 2.602]	[-0.357 ; 2.602]	[-0.181 ; 2.122]	[0.054 ; 2.122]
Information-processing score	[-3.396 ; 2.825]	[-0.370 ; 2.825]	[-0.142 ; 2.125]	[0.107 ; 2.125]
	[-3.437 ; 2.865]	[-0.390 ; 2.865]	[-0.202 ; 2.354]	[0.064 ; 2.354]
Number of observations	132676	132676	132676	132676
Number of clusters	5449	5449	5449	5449

Note: For each outcome variable and assumption combination, the first line shows the estimated bounds and the second line shows the 95% confidence intervals around the bounds. In column (3) and (4) the estimated bounds and confidence intervals are bias-corrected, following the method of Kreider and Pepper (2007) that was developed for MIV-based bounds. All confidence intervals are based on 1000 bootstrap replications, using the method that was developed by Imbens and Manski (2004). We take clustering at the school level into account.

Table 4.5. Results based on NO Assumptions, on the MTS assumption, on the MTS-MIV assumptions and on the MTR-MTS-MIV assumptions, in the schoolyear 1999/2000

	(1)	(2)	(3)	(4)
	NOA	MTS	MTS-MIV	MTR-MTS-MIV
Total Cito score	[-2.754 ; 2.151]	[-0.356 ; 2.151]	[-0.045 ; 1.810]	[0.136 ; 1.810]
	[-2.779 ; 2.176]	[-0.377 ; 2.176]	[-0.112 ; 1.961]	[0.081 ; 1.961]
Math score	[-2.870 ; 2.335]	[-0.267 ; 2.335]	[0.018 ; 1.983]	[0.122 ; 1.983]
	[-2.897 ; 2.362]	[-0.285 ; 2.362]	[-0.048 ; 2.144]	[0.066 ; 2.144]
Language score	[-3.317 ; 2.683]	[-0.322 ; 2.683]	[-0.010 ; 2.275]	[0.142 ; 2.275]
	[-3.348 ; 2.714]	[-0.342 ; 2.714]	[-0.087 ; 2.456]	[0.082 ; 2.456]
Information-processing score	[-3.558 ; 2.815]	[-0.387 ; 2.815]	[-0.105 ; 2.416]	[0.115 ; 2.416]
	[-3.593 ; 2.850]	[-0.407 ; 2.850]	[-0.166 ; 2.604]	[0.066 ; 2.604]
Number of observations	138079	138079	138079	138079
Number of clusters	5621	5621	5621	5621

Note: For each outcome variable and assumption combination, the first line shows the estimated bounds and the second line shows the 95% confidence intervals around the bounds. In column (3) and (4) the estimated bounds and confidence intervals are bias-corrected, following the method of Kreider and Pepper (2007) that was developed for MIV-based bounds. All confidence intervals are based on 1000 bootstrap replications, using the method that was developed by Imbens and Manski (2004). We take clustering at the school level into account.

Bibliography

Abeliovich, D., Leiberman, J. R., Teuerstein, I., and Levy, J. (1984). Prenatal sex diagnosis: Testosterone and FSH levels in mid-trimester amniotic fluids. *Prenatal Diagnosis*, 4:347–353.

Adams, B. N. (1999). *Handbook of Marriage and the Family*, chapter Cross-cultural and US kinship, pages 77–91.

Aghion, P. and Tirole, J. (1997). Formal and real authority in organizations. *The Journal of Political Economy*, 105(1):1–29.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, 97(4):543–569.

Almada, L., McCarthy, I. M., and Tchernis, R. (forthcoming). What can we learn about the effects of food stamps on obesity in the presence of misreporting? *American Journal of Agricultural Economics*.

Altonji, J. G., Cattan, S., and Ware, I. (2013). Identifying sibling influence on teenage substance use. *IFS Working Paper W13/04*.

Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *The American Economic Review*, 88:450–477.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Archer, J. (2005). An Edmonton journey. *Education Week*, 24(20):33–36.

Ariely, D. and Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*, 13(3):219–224.

Baizán, P., Aassve, A., and Billari, F. C. (2003). Cohabitation, marriage, and first birth: The interrelationship of family formation events in Spain. *European Journal of Population*, 19(2):147–169.

Baron-Cohen, S., Lutchmaya, S., and Knickmeyer, R. (2004). *Prenatal Testosterone in Mind - Amniotic Fluid Studies*. The MIT Press.

Benirschke, K., Burton, G. J., and Baergen, R. N. (2012). Multiple pregnancies. In *Pathology of the Human Placenta*. Springer.

Bertrand, M. (2010). *Handbook of Labor Economics*, chapter New Perspectives on Gender, pages 1545–1592. North Holland.

Björklund, A., Jäntti, M., and Lindquist, M. J. (2009). Family background and income during the rise of the welfare state: Brother correlations in income for Swedish men born 1932-1968. *Journal of Public Economics*, 93(5):671–680.

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*, 120(2):669–700.

Blanden, J. and Gregg, P. (2004). Family income and educational attainment: A review of approaches and evidence for Britain. *Oxford Review of Economic Policy*, 20(2):245–263.

Blundell, R., Gosling, A., Ichimura, H., and Meghir, C. (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, 75(2):323–363.

Bénabou, R., Kramarz, F., and Prost, C. (2009). The French zones d'éducation prioritaire: Much ado about nothing? *Economics of Education Review*, 28:345–356.

Bowles, H. R., Babcock, L., and Lai, L. (2007). Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organizational Behavior and Human Decision Processes*, 103:84–103.

Brim, O. G. (1958). Family structure and sex role learning by children: A further analysis of Helen Koch's data. *Sociometry*, 21:1–16.

Buser, T. and Peter, N. (2012). Multitasking. *Experimental Economics*, 15(4):641–655.

Butcher, K. F. and Case, A. (1994). The effect of sibling sex composition on women's education and earnings. *The Quarterly Journal of Economics*, 109(3):531–563.

Chen, S. H., Chen, Y.-C., and Liu, J.-T. (2013). The impact of family composition on human capital formation: Methodology and evidence. *Working paper*.

Clark, M. M., Tucker, L., and Galef, B. G. (1992). Stud males and dud males: intrauterine position effects on the reproductive success of male gerbils. *Animal Behaviour*, 43:215–221.

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

Cohen-Bendahan, C. C. C., van Goozen, S. H. M., Buitelaar, J. K., and Cohen-Kettenis, P. T. (2005b). Maternal serum steroid levels are unrelated to fetal sex: A study in twin pregnancies. *Twin Research and Human Genetics*, 8(2):173–177.

Conley, D. and Glauber, R. (2006). Parental educational investment and children's academic risk: Estimates of the impact of sibship size and birth order from exogenous variation in fertility. *The Journal of Human Resources*, 41(4):722–737.

Cools, S. and Kaldager Hart, R. (2015). The effect of childhood family size on fertility in adulthood. *Discussion papers No. 802 Statistics Norway.*

Copen, C. E., Daniels, K., Vespa, J., and Mosher, W. D. (2012). First marriages in the United States: Data from the 2006-2010 National Survey of Family Growth. In *National health statistics reports; no 49.* Hyattsville, MD: National Center for Health Statistics.

Coviello, D., Ichino, A., and Persico, N. (2011). Don't spread yourself too thin: The impact of task juggling on workers' speed of job completion. *Working paper.*

Criss, B. R. (2006). Gender differences in multitasking. *National Undergraduate Research Clearinghouse*, 9.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47:448–474.

Dahl, G. B., Løken, K. V., and Mogstad, M. (2014). Peer effects in program participation. *American Economic Review*, 104:2049–2074.

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

Danielsbacka, M., Tanskanen, A. O., Jokela, M., and Rotkirch, A. (2011). Grandparental child care in Europe: Evidence for preferential investment in more certain kin. *Evolutionary Psychology*, 9(1):3–24.

De Haan, M. (2011). The effect of parents' schooling on child's schooling: A nonparametric bounds analysis. *Journal of Labor Economics*, 29(4):859–892.

De Haan, M. (forthcoming). The effect of additional funds for low-ability pupils: A nonparametric bounds analysis. *Economic Journal*.

De Haan, M. and Leuven, E. (2016). Head Start and the distribution of long term education and labor market outcomes. *Working paper*.

Delaney, J. G. (1995). The development of school-based management in the Edmonton public school district. *Working paper*.

DfE (2011). Consultation on school funding reform: Proposals for a fairer system. *Department for Education*. Available online at https://www.education.gov.uk/consultations/downloadableDocs/July 2011 Consultation on School Funding Reform FINAL.pdf.

Dobbelsteen, S., Levin, J., and Oosterbeek, H. (2002). The causal effect of class size on scholastic achievement: distinguishing the pure class size effect from the effect of changes in class composition. *Oxford Bulleting of Economics and Statistics*, 64(17):17–38.

Drago, R. and Garvey, G. T. (1998). Incentives for helping on the job: Theory and evidence. *Journal of Labor Economics*, 16(1):1–25.

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, 58(5):1–113.

Dux, P. E., Ivanoff, J., Asplund, C. L., and Marios, R. (2006). Isolation of central bottleneck of information processing with time-resolved fMRI. *Neuron*, 52(6):1109–1120.

Dux, P. E., Tombu, M. N., Harrison, S., Rogers, B. P., Tong, F., and Marois, R. (2009). Training improves multitasking performance by increasing the speed of information processing in human prefrontal cortex. *Neuron*, 63(1):127–138.

Elston, R. C., Olson, J. M., and Palmer, L. (2002). *Biostatistical Genetics and Genetic Epidemiology*. John Wiley and Sons.

Even, M. D., Dhar, M. G., and vom Saal, F. S. (1992). Transport of steroids between fetuses via amniotic fluid in relation to the intrauterine position phenomenon in rats. *Journal of Reproduction and Fertility*, 96:709–716.

Falk, A. and Kosfeld, M. (2006). The hidden costs of control. *The American Economic Review*, 96(5):1611–1630.

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

Fisher, H. (1999). *The First Sex: The Natural Talents of Women and How They Are Changing the World*. Random House.

Fisher-Thompson, D. (1993). Adult toy purchases for children: Factors affecting sex-typed toy selection. *Journal of Applied Developmental Psychology*, 14:385–406.

Fordham Institute (2006). Fund the child: Tackling inequity and antiquity in school finance. *Washington, DC: Author.*

Friebel, G. and Yilmaz, L. (2010). The hidden costs of multi-tasking. *Working paper.*

Fudenberg, D. and Levine, D. K. (2006). A dual-self model of impulse control. *The American Economic Review*, 96(5):1449–1476.

Furtick, K. and Snell, L. (2013). Weighted Student Formula yearbook. *Reason Foundation*, 12/2013(426).

Gerfin, M. and Schellhorn, M. (2006). Nonparametric bounds on the effect of deductibles in health care insurance on doctor visits - Swiss evidence. *Health Economics*, 15(9):1011–1020.

Gibbons, S., McNally, S., and Viarengo, M. (2012). Does additional spending help urban schools? An evaluation using boundary discontinuities. *IZA Discussion Paper No. 6281.*

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.

Giustinelli, P. (2011). Non-parametric bounds on quantiles under monotonicity assumptions: With an application to the Italian education returns. *Journal of Applied Econometrics*, 26(5):783–824.

Glass, A. and Klein, T. (1981). Changes in maternal serum total and free androgen levels in early pregnancy: lack of correlation with fetal sex. *American Journal of Obstetrics and Gynecology*, 140:656–660.

Glucksberg, S. (1962). The influence of strength of drive on functional fixedness and perceptual recognition. *Journal of Experimental Psychology*, 63(1):36–41.

Gonzalez, V. M. and Mark, G. (2004). Constant, constant multi-tasking craziness: Managing multiple working spheres. *CHI Letters*, 6(1):113–120.

Gray, R. H., Simpson, J. L., Bitto, A. C., Queenan, J. T., Li, C., Kambic, R. T., Perez, A., Mena, P., Barbato, M., Stevenson, W., and Jennings, V. (1998). Sex ratio associated with timing of insemination and length of the follicular phase in planned and unplanned pregnancies during the use of natural family planning. *Human Reproduction*, 13(5):1397–1400.

Gundersen, C., Kreider, B., and Pepper, J. (2012). The impact of the National School Lunch Program on child health: A nonparametric bounds analysis. *Journal of Econometrics*, 166(1):79–91.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal*, 113(485):F64–F98.

Hauser, R. M. and Kuo, H.-H. D. (1998). Does the gender composition of sibships affect women's educational attainment? *The Journal of Human Resources*, 33(3):644–657.

Havel, M. A. (2004). Gender differences in multitasking abilities. *National Undergraduate Research Clearinghouse*, 7.

Haveman, R. and Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature*, 33:1829–1878.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Hægelanda, T., Raaum, O., and Salvanes, K. G. (2012). Pennies from heaven? Using exogenous tax variation to identify effects of school resources on pupil achievement. *Economics of Education Review*, 31(5):601–614.

Hof, S. (2014). Does private tutoring work? The effectiveness of private tutoring: A nonparametric bounds analysis. *Education Economics*, 22(4):347–366.

Holmlund, H., Lindahl, M., and Plug, E. (2011). The causal effect of parents' schooling on children's schooling: A comparison of estimation methods. *Journal of Economic Literature*, 49(3):615–651.

Holmlund, H., McNally, S., and Viarengo, M. (2010). Does money matter for schools? *Economics of Education Review*, 29:1154–1164.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics & Organization*, 7:24–52.

Husby, H., NV, H., A, G., SG, T., K, K., and H, G. (1991). Zygosity, placental membranes and Weinberg's rule in a Danish consecutive twin series. *Acta Geneticae Medicae et Gemellologiae*, 40:147–152.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Jackson, C. K., Johnson, R. C., and Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics*, 131(1):157–218.

James, W. H. (1979). Is Weinberg's differential rule valid? *Acta Geneticae Medicae et Gemellologiae*, 28(1):69–71.

Jayachandran, S. and Pande, R. (2015). Why are Indian children so short? *NBER Working Paper Series*, Working Paper 21036.

Joensen, J. S. and Nielsen, H. S. (2015). Spillovers in educational choice. *Working paper*.

Kaestner, R. (1997). Are brothers really better? Sibling sex composition and educational achievement revisited. *The Journal of Human Resources*, 32(2):250–284.

Kahneman, D. and Tversky, A. (1984). Choices, values and frames. *American Psychologist*, 39(4):341–350.

Kalenkoski, C. M. and Foster, G. (2010). The multitasking of household production. *IZA Discussion Paper No. 4845*.

Keim, S., Klärner, A., and Bernardi, L. (2009). Who is relevant? Exploring fertility relevant social networks. *MPIDR Working paper WP 2009-001*.

Klebanov, P. K., Brooks-Gunn, J., McCarton, C., and McCormick, M. C. (1998). The contribution of neighborhood and family income to developmental test scores over the first three years of life. *Child Development*, 69(5):1420–1436.

Koch, H. L. (1955). The relation of certain family constellation characteristics and the attitudes of children toward adults. *Child Development*, 26:13–40.

Kolk, M. (2015). The causal effect of an additional sibling on completed fertility: An estimation of intergenerational fertility correlations by looking at siblings of twins. *Demographic Research*, 32:1409–1420.

Kreider, B. and Hill, S. C. (2009). Partially identifying treatment effects with an application to covering the uninsured. *Journal of Human Resources*, 44(2):409–449.

Kreider, B., Manski, R. J., Moeller, J., and Pepper, J. (2013). Estimating the effect of dental insurance on the use of dental services when true coverage is unobserved. *Working paper.*

Kreider, B. and Pepper, J. (2008). Inferring disability status from corrupt data. *Journal of Applied Econometrics*, 23(3):329–349.

Kreider, B. and Pepper, J. V. (2007). Disability and employment. *Journal of the American Statistical Association*, 102(478):432–441.

Kreider, B., Pepper, J. V., and Roy, M. (2016). Identifying the effects of WIC on food insecurity among infants and children. *Southern Economic Journal*, 82(4):1106–1122.

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal*, 113(485):F34–F63.

Kuegler, A. (2009). A curse of comparison? Evidence on reference groups for relative income concerns. *Policy Research Working Paper 4820.*

Ladd, H. F. and Fiske, E. B. (2011). Weighted Student Funding in the Netherlands: A model for the U.S.? *Journal of Policy Analysis and Management*, 30(3):470–498.

Lazear, E. P. and Gibbs, M. (2009). *Personnel Economics in Practice.* Wiley.

Leuven, E., Lindahl, M., Oosterbeek, H., and Webbink, D. (2007). The effect of extra funding for disadvantaged pupils on achievement. *The Review of Economics and Statistics*, 89(4):721–736.

Lichtenstein, P., Faire, U. D., Floderus, B., Svartengren, M., Svedberg, P., and Pedersen, N. L. (2002). The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine*, 252:184–205.

Lindbeck, A. and Snower, D. J. (2000). Multitask learning and the reorganization of work: From Tayloristic to holistic organization. *Journal of Labor Economics*, 18(3):353–376.

Machin, S., McNally, S., and Meghir, C. (2010). Resources and standards in urban schools. *Journal of Human Capital*, 4(4):365–393.

Manski, C. F. (1989). Anatomy of the selection problem. *The Journal of Human Resources*, 24(3):343–360.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, 65(6):1311–1334.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010.

Manski, C. F. and Pepper, J. V. (2009). More on monotone instrumental variables. *Econometrics Journal*, 12:S200–S216.

Manski, C. F. and Pepper, J. V. (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, 29(1):123–141.

Mariotti, M. and Meinecke, J. (2015). Partial identification and bound estimation of the average treatment effect of education on earnings for South Africa. *Oxford Bulletin of Economics and Statistics*, 77(2):210–233.

McHale, S. M. and Crouter, A. C. (1996). The family contexts of children's sibling relationships. In Brody, G. H., editor, *Sibling relationships: Their causes and consequences*, pages 173–195. Norwood, NJ: Ablex.

Meulenberg, P. M. M. and Hofman, J. A. (1991). Maternal testosterone and fetal sex. *The Journal of Steroid Biochemistry and Molecular Biology*, 39(1):51–54.

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

Monsell, S. (2003). Task switching. *TRENDS in Cognitive Sciences*, 7(3):134–140.

Morduch, J. (2000). Sibling rivalry in Africa. *The American Economic Review*, 90(2):405–409. Papers and Proceedings of the One Hundred Twelfth Annual Meeting of the American Economic Association.

Mueller, G. and Plug, E. (2006). Estimating the effect of personality on male and female earnings. *Industrial and Labor Relations Review*, 60:3–22.

Mulder, L. and Van der Werf, G. (1997). Implementation and effects of the Dutch Educational Priority Policy: Results of four years of evaluation studies. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 3(4):317–339.

Nelson, A. (2005). Children's toy collections in Sweden - a less gender-typed country? *Sex Roles*, 52:93–102.

Nicoletti, C., Peracchi, F., and Foliano, F. (2011). Estimating income poverty in the presence of missing data and measurement error. *Journal of Business and Economic Statistics*, 29(1):61–72.

Nicoletti, C. and Rabe, B. (2014). Sibling spillover effects in school achievement. *IZA Discussion Paper No. 8615.*

OECD (2012). Equity and quality in education: Supporting disadvantaged students and schools. *OECD Publishing.*

Ohinata, A. and Van Ours, J. C. (2012). Young immigrant children and their educational attainment. *Economics Letters*, 116(3):288–290.

Ohinata, A. and Van Ours, J. C. (2013). How immigrant children affect the academic achievement of native Dutch children. *The Economic Journal.*

Parish, W. L. and Willis, R. J. (1993). Daughters, education, and family budgets: Taiwan experiences. *The Journal of Human Resources*, 28(4):863–898. Special Issue: Symposium on Investments in Women's Human Capital and Development.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, 116(2):220–244.

Pease, A. and Pease, B. (2001). *Why Men Don't Listen and Women Can't Read Maps: How We're Different and What to Do About It.* Three Rivers Press.

Pease, A. and Pease, B. (2003). *Why Men Can Only Do One Thing at a Time Women Never Stop Talking.* Orion.

Peter, N., Lundborg, P., and Webbink, D. (2015). The effect of a sibling's gender on earnings, education and family formation. *IZA Discussion Paper No. 9128.*

Pettersson-Lidbom, P., Thoursie, P. S., and Vlachos, J. (2008). Does child gender affect sibling and family outcomes? *Unpublished draft.*

Phillips, D. I. W. (1993). Twin studies in medical research: can they tell us whether diseases are genetically determined? *The Lancet*, 341:1008–1009.

Pollet, T. V., Nelissen, M., and Nettle, D. (2009). Lineage based differences in grandparental investment: evidence from a large British cohort study. *Journal of Biosocial Science*, 41:355–379.

Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B. (2010). *Biology - Campbell, 9th Edition.* Benjamin Cummings.

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.

Richey, J. (2014). The effect of youth labor market experience on adult earnings. *Journal of Economic Development*, 39(1):47–61.

Rodeck, C. H., Gill, D., Rosenberg, D. A., and Collins, W. P. (1985). Testosterone levels in midtrimester maternal and fetal plasma and amniotic fluid. *Prenatal Diagnosis*, 5:175–181.

Roeleveld, J., Driessen, G., Ledoux, G., Cuppen, J., and Meijer, J. (2011). Doelgroepleerlingen in het basisonderwijs: Historische ontwikkeling en actuele situatie. *Amsterdam: Kohnstamm Instituut*, Rapport 857.

Rubinstein, J. S., Meyer, D. E., and Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):763–797.

Rust, J., Golombok, S., Hines, M., Johnston, K., and Golding, J. (2000). The role of brothers and sisters in the gender development of preschool children. *Journal of Experimental Child Psychology*, 77:292–303.

Ryan, B. and Vandenbergh, J. (2002). Intrauterine position effects. *Neuroscience and Biobehavioral Reviews*, 26:665–678.

Schottner, A. (2007). Relational contracts, multitasking, and job design. *The Journal of Law, Economics, & Organization*, 24(1):138–162.

Shiv, B. and Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*, 26:278–292.

Singer, E., van Hoewyk, J., and Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64:171–188.

Åslund, O. and Grönqvist, H. (2007). Family size and child outcomes: Is there really no trade-off? *Working paper.*

Åslund, O. and Grönqvist, H. (2010). Family size and child outcomes: Is there really no trade-off? *Labour Economics*, 17:130–139.

Spence, K. W., Farber, I. E., and McFann, H. H. (1956). The relation of anxiety (drive) level to performance in competitional and noncompetitional paired-associates learning. *Journal of Experimental Psychology*, 52:296–305.

Stoneman, Z., Brody, G. H., and MacKinnon, C. E. (1986). Same-sex and cross-sex siblings: Activity choices, roles, behavior, and gender stereotypes. *Sex Roles*, 15:495–511.

Säve-Söderbergh, J. (2007). Are women asking for low wages? Gender differences in wage bargaining strategies and ensuing bargaining success. *Working paper*.

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*, 60:713–722.

Toriola, A. T., Vääräsmäki, M., Lehtinen, M., Zeleniuch-Jacquotte, A., Lundin, E., Rodgers, K.-G., Lakso, H.-A., Chen, T., Schock, H., Hallmans, G., Pukkala, E., Toniolo, P., Grankvist, K., Surcel, H.-M., and Lukanova, A. (2011). Determinants of maternal sex steroids during the first half of pregnancy. *Obsterics and Gynecology*, 118(5).

Tucker, C. J., Updegraff, K., and Baril, M. E. (2010). Who's the boss? Patterns of control in adolescents' sibling relationships. *Family Relations*, 59:520–532.

UNECE (2012). Mean age at first marriage by sex. In *UNECE Statistical Division Database*. United Nations Economic Commission for Europe.

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*, 46:663–669.

Van der Klaauw, W. (2008). Breaking the link between poverty and low student achievement: An evaluation of Title I. *Journal of Econometrics*, 142:731–756.

Vlietinck, R., C, D., R, D., den Berghe H, V., and M, T. (1988). The validity of Weinberg's rule in the East Flanders Prospective Twin Survey (EFPTS). *Acta Geneticae Medicae et Gemellologiae*, 37:137–141.

vom Saal, F. S. and Bronson, F. H. (1980). Sexual characteristics of adult female mice are correlated with their blood testosterone levels during prenatal development. *Science*, 208:597–599.

vom Saal, F. S., Grant, W. M., McMullen, C. W., and Laves, K. S. (1983). High fetal estrogen concentrations: Correlation with increased adult sexual activity and decreased aggression in male mice. *Science*, 220:1306–1309.

Wilcox, A. J., Weinberg, C. R., and Baird, D. D. (1995). Timing of sexual intercourse in relation to ovulation. *The New England Journal of Medicine*, 333(23):1517–1521.

Zábojník, J. (2002). Centralized and decentralized decision making in organizations. *Journal of Labor Economics*, 20(1):1–22.

Summary (in English)

The title of this thesis is “Essays in Empirical Microeconomics”. It contains three empirical studies that investigate factors that could affect individuals’ labor market, family and educational outcomes. Chapter 2 focuses on scheduling as a potential determinant of individuals’ productivity. Chapter 3 examines the role of a family factor on children’s long term outcomes. In particular, it examines how the gender of a sibling affects earnings, education and family formation. Chapter 4 looks at the impact of school resources on the performance of students in a Weighted Student Funding setting. In this summary I give a brief overview of each of these studies in turn.

Chapter 2 examines four questions. First, it examines how multitasking affects productivity. Second, it examines whether individuals optimally choose their degree of multitasking or whether they perform better under an externally imposed schedule. Third, it examines whether there are indeed gender differences in the effect of multitasking on productivity, as it is often hypothesized. Finally, it examines whether there are gender differences in the propensity to multitask. We examine these research questions empirically by conducting a laboratory experiment in which subjects are randomly allocated to different work schedules. We examine performance under three different schedules: one where subjects perform two tasks sequentially, one where subjects are forced to multitask, and one where subjects can freely organize their work. The amount of time spent on each task is identical in each treatment. Thus, performance differences between treatments measure the productivity effect of the different schedules.

We find that subjects who are forced to multitask perform significantly worse than those forced to work sequentially. Surprisingly, subjects who can freely organize their own schedule also perform significantly worse. These results suggest that scheduling is a significant determinant of productivity. Finally, our results do not support the stereotype that women are better at multitasking. Women suffer as much as men when forced to multitask and are actually less inclined to multitask when being free to choose.

Chapter 3 studies how the gender of a sibling affects earnings, education and family formation. Identification is complicated by parental preferences: if parents prefer certain sex compositions over others, children’s gender affects not only the outcomes

of other children but also the existence of potential additional children. We apply an empirical approach that circumvents this problem. Using a sample of dizygotic (i.e. non-identical) twins, we compare men (women) with co-twin brothers to men (women) with co-twin sisters. In these cases, the two children are being born at the same time, so parents cannot make decisions about one twin based on the gender of the other twin.

We find that the gender of the sibling influences both men and women, but in a different way. Men with a co-twin brother earn more and are more likely to get married and have children than men with a co-twin sister. Women with a co-twin sister obtain lower education and give birth earlier than women with a co-twin brother. Our analysis shows that the most likely explanation for these findings is that siblings affect each other via various social mechanisms.

Chapter 4 focuses on a Dutch policy measure that provides extra funding to schools for students with a disadvantaged family background. The policy allocates funding to primary schools based on a national formula that includes extra weights for disadvantaged students. We use a nonparametric method to estimate upper and lower bounds on the effect of the extra resources on test scores. We start the analysis by calculating the worst-case bounds that can be obtained without imposing assumptions, and then we layer various nonparametric assumptions to tighten the bounds. We make use of three types of assumptions: Monotone Treatment Selection, Monotone Instrumental Variable and Monotone Treatment Response. For the MIV bounds, we use average neighborhood income, thereby assuming a non-negative relationship between mean potential test scores and average income in the neighborhood.

The MTS-MIV bounds indicate that the extra funding has a positive impact on math scores. The bounds around the impact on language scores, information-processing scores and total scores also allow for large positive impacts, but they also allow for a more moderate negative impact. We subsequently show that adding the MTR assumption tightens the bounds such that all lower bounds are significantly positive, but we caution against drawing firm conclusions from MTR-based bounds as they rely on stronger assumptions.

Clarification to co-authored papers:

Chapter 2
Title: Multitasking
Co-author: Thomas Buser
Publication: Experimental Economics, 15(4):641-655
Contribution of the doctoral candidate: She is one of the main contributors of the study. She initiated the research project, was to a large extent responsible for the design and implementation of the experiment, and did a substantial part of the data analysis and the writing.

Chapter 3
Title: The effect of a sibling's gender on earnings, education and family formation
Co-authors: Petter Lundborg and Dinand Webbink
Publication: IZA Discussion Paper No. 9128
Contribution of the doctoral candidate: She is the main contributor of the study. The creative part of the study (the identification strategy and other main points) relies heavily on her ideas. She conducted the empirical analysis herself and was also responsible for the writing.

Samenvatting (Summary in Dutch)

Mijn proefschrift is getiteld ”Essays in Empirical Microeconomics”. Het bevat drie empirische papers waarin factoren onderzocht worden die effect kunnen hebben op de arbeidsmarkt, familie en onderwijsprestaties. Hoofdstuk 2 concentreert zich op planning als een potentiële determinant van de individuele arbeidsproductiviteit. Hoofdstuk 3 onderzoekt de rol van het geslacht van een kind binnen een gezin op lange termijn uitkomsten voor de andere kinderen. In het bijzonder bekijken we of het geslacht van een kind invloed heeft op inkomsten, genoten onderwijs en familievorming van het andere kind. In hoofdstuk 4 onderzoeken wij de invloed van beschikbare schoolmiddelen op leerling-prestaties wanneer er een gewichtenregeling is. In deze samenvatting geef ik een kort overzicht van deze onderzoeken.

In hoofdstuk 2 worden vier vragen beantwoord. Allereerst, de invloed van multitasken op productiviteit. Ten tweede, de vraag of individuen het optimale niveau van multitasken kunnen bepalen, of dat ze beter presteren onder een extern opgelegd regime. Ten derde, onderzoeken wij of er inderdaad gender-verschillen zijn bij de invloed van multitasken op productiviteit, zoals vaak wordt gesteld. Als laatste onderzoeken wij of er gender-verschillen zijn in de neiging om te multitasken. We onderzoeken deze onderzoeksvragen empirisch met een gerandomiseerd laboratorium experiment. De deelnemers zijn willekeurig toegewezen aan verschillende werk schema's. We bekijken de prestaties onder drie verschillende regimes: één waarbij de deelnemers twee taken achter elkaar volbrengen; één waarbij de deelnemers gedwongen worden te multitasken en één waarbij de deelnemers zelf vrij zijn het regime te bepalen.

De hoeveelheid tijd die aan elke taak wordt besteed is gelijk in elke behandeling. Als gevolg, kunnen prestatie-verschillen veroorzaakt door de verschillende behandelingen, worden geïnterpreteerd als het effect van de behandelingen op productiviteitsverschillen.

Wij vinden dat deelnemers die tot multitasken gedwongen worden, significant slechter presteren dan wanneer ze het werk achter elkaar moeten doen. Verrassend genoeg, presteren ook de deelnemers die mogen kiezen tussen multitasken of sequentieel werken, significant slechter. Deze uitkomsten suggereren dat planning een belangrijke determinant is van productiviteit. Als laatste, geven onze resultaten geen steun voor het stereotype beeld dat vrouwen beter zijn in multitasken. Vrouwen presteren net zo slecht als mannen wanneer ze gedwongen worden om te multitaksen. Vrouwen zijn bovendien minder geneigd om voor multitasken te kiezen, wanneer ze de keuze krijgen.

In hoofdstuk 3 onderzoeken wij of het geslacht van een kind, invloed heeft op inkomsten, onderwijs en familievorming van het andere kind in de familie. Identificeren van dit effect is gecompliceerd door de voorkeuren van ouders voor het geslacht van hun kinderen. Wanneer ouders een voorkeur hebben voor een bepaalde gender-samenstelling van het gezin boven andere samenstellingen, dan beïnvloed het geslacht van een kind niet alleen de uitkomsten van broers en zussen, maar zelfs het bestaan van extra broers of zussen.

We passen een empirische strategie toe die dit probleem omzeilt. We gebruiken een steekproef van dizygotic (twee-eiige) tweelingen, waarin we mannen (vrouwen) met een tweelingbroer vergelijken met mannen (vrouwen) met een tweelingzus. In deze gevallen zijn de kinderen geboren op hetzelfde moment, zodat de ouders geen beslissing hebben kunnen nemen over het ene kind, gebaseerd op het geslacht van het andere kind.

Wij vinden dat het geslacht van een kind invloed heeft op zowel mannen als vrouwen, maar wel op een andere manier. Mannen met een tweelingbroer verdienen meer, hebben een grotere kans te trouwen en kinderen te krijgen, dan mannen met een tweelingzus. Vrouwen met een tweelingzus bereiken een lager onderwijsniveau en hebben jonger kinderen, dan vrouwen met een tweelingbroer. Onze analyse laat zien dat de meest waarschijnlijke verklaring voor onze bevindingen is dat kinderen elkaar beïnvloeden via verschillende kanalen.

Hoofdstuk 4 concentreert zich op een Nederlands beleid waarin extra middelen beschikbaar zijn voor leerlingen die vanwege hun familieachtergrond extra hulp nodig hebben. De middelen komen ter beschikking van basisscholen gebaseerd op de gewichtenregeling. Hierin bepalen de kenmerken van de familie van de leerling het gewicht en daarmee de extra middelen voor de school.

Wij gebruiken een niet-parametrische methode om de bovenste en de onderste grens te schatten op het effect van de extra middelen op de testscores van de leerlingen. Wij beginnen met de analyse door de grenzen te berekenen die verkregen worden zonder beperkingen op te leggen, en daarna voegen we stap voor stap verschillen niet-parametrische aannames toe om de grenzen dichter bij elkaar te brengen. We gebruiken drie typen aannames: monotone behandeling selectie (MTS), monotone instrumentele variatie (MIV) en monotone behandeling reactie (MTR). Bij de MIV bandbreedte, gebruiken we een gemiddeld buurtinkomen, waarbij we aannemen dat er een niet-negatieve relatie is tussen de gemiddelde potentiële testscores en het gemiddelde inkomen in een buurt.

De MTS-MIV bandbreedtes geven aan dat extra middelen een positief effect hebben op wiskunde scores. De grenzen van het effect op taalprestaties, studievaardigheden (informatieverwerking) en totale prestaties, lijken grote positieve effecten toe te staan, maar deze staan ook een meer bescheiden negatief effect toe, zodat de richting van het effect onduidelijk is. Wij laten vervolgens zien dat het toevoegen van de MTR aanname, de grenzen vernauwt zodat alle ondergrenzen significant positief zijn, maar wij waarschuwen tegen het trekken van stevige conclusies, omdat de MTR-grenzen van sterke aannames afhankelijk zijn.

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:

608 Ł.M. MARĆ, *The Impact of Aid on Total Government Expenditures*

609 C. LI, *Hitchhiking on the Road of Decision Making under Uncertainty*

610 L. ROSENDAHL HUBER, *Entrepreneurship, Teams and Sustainability: a Series of Field Experiments*

611 X. YANG, *Essays on High Frequency Financial Econometrics*

612 A.H. VAN DER WEIJDE, *The Industrial Organization of Transport Markets: Modeling pricing, Investment and Regulation in Rail and Road Networks*

613 H.E. SILVA MONTALVA, *Airport Pricing Policies: Airline Conduct, Price Discrimination, Dynamic Congestion and Network Effects.*

614 C. DIETZ, *Hierarchies, Communication and Restricted Cooperation in Cooperative Games*

615 M.A. ZOICAN, *Financial System Architecture and Intermediation Quality*

616 G. ZHU, *Three Essays in Empirical Corporate Finance*

617 M. PLEUS, *Implementations of Tests on the Exogeneity of Selected Variables and their Performance in Practice*

618 B. VAN LEEUWEN, *Cooperation, Networks and Emotions: Three Essays in Behavioral Economics*

619 A.G. KOPÁNYI-PEUKER, *Endogeneity Matters: Essays on Cooperation and Coordination*

620 X. WANG, *Time Varying Risk Premium and Limited Participation in Financial Markets*

621 L.A. GORNICKA, *Regulating Financial Markets: Costs and Trade-offs*

622 A. KAMM, *Political Actors playing games: Theory and Experiments*

623 S. VAN DEN HAUWE, *Topics in Applied Macroeconometrics*

624 F.U. BRÄUNING, *Interbank Lending Relationships, Financial Crises and Monetary Policy*

625 J.J. DE VRIES, *Estimation of Alonso's Theory of Movements for Commuting*

626 M. POPŁAWSKA, *Essays on Insurance and Health Economics*

627 X. CAI, *Essays in Labor and Product Market Search*

628 L. ZHAO, *Making Real Options Credible: Incomplete Markets, Dynamics, and Model Ambiguity*

629 K. BEL, *Multivariate Extensions to Discrete Choice Modeling*

630 Y. ZENG, *Topics in Trans-boundary River sharing Problems and Economic Theory*

631 M.G. WEBER, *Behavioral Economics and the Public Sector*

632 E. CZIBOR, *Heterogeneity in Response to Incentives: Evidence from Field Data*

633 A. JUODIS, *Essays in Panel Data Modelling*

634 F. ZHOU, *Essays on Mismeasurement and Misallocation on Transition Economies*

635 P. MULLER, *Labor Market Policies and Job Search*

636 N. KETEL, *Empirical Studies in Labor and Education Economics*

637 T.E. YENILMEZ, *Three Essays in International Trade and Development*

638 L.P. DE BRUIJN, *Essays on Forecasting and Latent Values*

639 S. VRIEND, *Profiling, Auditing and Public Policy: Applications in Labor and Health Economics*

640 M.L. ERGUN, *Fat Tails in Financial Markets*

641 T. HOMAR, *Intervention in Systemic Banking Crises*

642 R. LIT, *Time Varying Parameter Models for Discrete Valued Time Series*

643 R.H. KLEIJN, *Essays on Bayesian Model Averaging using Economic Time Series*

644 S. MUNS, *Essays on Systemic Risk*

645 B.M. SADABA, *Essays on the Empirics of International Financial Markets*

646 H. KOC, *Essays on Preventive Care and Health Behaviors*

647 V.V.M. MISHEVA, *The Long Run Effects of a Bad Start*

648 W. LI, *Essays on Empirical Monetary Policy*

649 J.P. HUANG, *Topics on Social and Economic Networks*

650 K.A. RYSZKA, *Resource Extraction and the Green Paradox: Accounting for Political Economy Issues and Climate Policies in a Heterogeneous World*

651 J.R. ZWEERINK, *Retirement Decisions, Job Loss and Mortality*

652 M. K. KAGAN, *Issues in Climate Change Economics: Uncertainty, Renewable Energy Innovation and Fossil Fuel Scarcity*

653 T.V. WANG, *The Rich Domain of Decision Making Explored: The Non-Triviality of the Choosing Process*

654 D.A.R. BONAM, *The Curse of Sovereign Debt and Implications for Fiscal Policy*

655 Z. SHARIF, *Essays on Strategic Communication*

656 B. RAVESTEIJN, *Measuring the Impact of Public Policies on Socioeconomic Disparities in Health*

657 M. KOUDSTAAL, *Common Wisdom Versus Facts; How Entrepreneurs Differ in Their Behavioral Traits From Others*

The empirical studies in this thesis investigate various factors that could affect individuals' labor market, family formation and educational outcomes. Chapter 2 focuses on scheduling as a potential determinant of individuals' productivity. Chapter 3 looks at the role of a family factor on children's long term outcomes. In particular, it examines how the gender of a sibling affects individuals' earnings, education and family formation. Chapter 4 examines the impact of school resources on the performance of students in a Weighted Student Funding setting. Each chapter applies a different method to identify causal effects. The first study uses a lab experiment, the second study exploits a natural experiment, and the third study uses a nonparametric bound method. While the three studies use different methods, a common feature is that they all try to deal with some kind of a selection problem to get to the causal effect.

Noemi Peter completed the Tinbergen Institute's Master of Philosophy in Economics program in 2010 (Cum Laude). She continued her studies as a PhD student under the supervision of Hessel Oosterbeek at the University of Amsterdam. From August 2016 onward she works as an Assistant Professor at the University of Groningen.