Essays in Labor Economics and the Economics of Education

DISSERTATION
of the University of St. Gallen,
School of Management,
Economics, Law, Social Sciences
and International Affairs
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Petra Maria Thiemann

from
Germany

Approved on the application of

Prof. Dr. Michael Lechner

and

Prof. Bryan S. Graham, PhD

Dissertation no. 4438
Zeitdruck GbR

The University of St. Gallen, School of Management, Economics, Law, Social Sciences and International Affairs hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St. Gallen, May 20, 2015

The President:

Prof. Dr. Thomas Bieger

To my parents

Acknowledgments

While I was writing this thesis, I received support and encouragement from a variety of people. In fact, before I started this thesis, I did not expect that completing a PhD could be such a social endeavor. But luckily, it was. I was fortunate to meet colleagues and friends who nourished me intellectually, and who carried me through the inevitable ups and downs.

First and foremost, I would like to thank my adviser Michael Lechner, who introduced me to the art of causal analysis. I will always be grateful for his openness and accessibility, his comments and attention to details. My work benefited greatly from his guidance. I would also like to thank him for his early trust in my motivation to conduct empirical research. He hired me as a student researcher at a time when I did not have any appreciable experience in the field.

I am indebted to my co-adviser Bryan S. Graham, who invited me to spend a year at UC Berkeley to deepen my understanding of the analysis of social interactions. I am grateful for his generosity with the time that he spent on sharing his knowledge, reading, discussing and commenting on my work, helping and advising me during my job search process, and motivating and supporting me in my development as a researcher. I am especially happy that he came to St. Gallen to serve on my dissertation committee.

Moreover, I would like to thank Beatrix Eugster, the third member of my dissertation committee, for her useful comments on the last two chapters of this thesis.

A special thanks goes to my co-authors Monika Bütler, Eva Deuchert, Christina Felfe, Sharon Pfister, Darjusch Tafreschi, and Stefan Staubli. Their knowledge and skills, their guidance, patience, and eagerness to discuss were valuable to me. Working side by side with them shaped me in important ways.

I also thank my co-workers at the Swiss Institute for Empiricial Economic Research for many fruitful discussions and interactions: Stefanie Behncke, Charlotte Cabane, Hans Fricke, Martin Huber, Lukas Inderbitzin, Katharina Jaronicki, Winfried Königer, Giovanni Mellace, Jonathan Schulz, Jan Richard Schumacher, Benjamin Schünemann, Andreas Steinmayr, Rudi Stracke, Tony Strittmatter, Uwe Sunde, Christian Thöni, Thomas Vischer, Conny Wunsch, and Susanne Zopf. They contributed to an intellectually stimulating work environment. They also filled my workdays as well as our time off together with joy.

I was fortunate to meet many people who accompanied me during my studies and student life since I started my Master's in St. Gallen, and who became good friends over time. In particular, I would like to thank Nina Büttner, Maria Dieterle, Daniel Kienzler, Jan-Aaron Klaassen, Elena Manresa, Philip Schuster, Alexa Tiemann, and Martin Wermelinger. I am grateful for their company and for learning from them, and I am glad that we could share important moments on our paths.

Finally, a huge thanks goes to my family. I dedicate this thesis to my parents, who have always supported me unconditionally and in all conceivable ways. I thank my sister and her husband, Anna and Tobias Lange, as well as my brother Christian Thiemann for their steady care and attention, no matter where I am.

Los Angeles, August 2015
Petra Thiemann

Contents

List of Figures

List of Tables

Summary

This thesis combines five essays in the fields of Labor Economics and the Economics of Education. The goal of the thesis is to understand the factors that determine individuals' choices with respect to their educational attainment and their labor supply. The thesis is motivated by the notion that policies at different institutional levels (e.g., at the university or at the government level) can influence these choices to some extent.

The first two chapters examine the role of peer groups for student outcomes in post-secondary education. Many university entrants rely on friends and study partners as sources of information and support. To determine the effect of peer group composition on academic achievement, I exploit random assignment to orientation week groups at the University of St. Gallen. Chapter 1 examines the effect of the composition of these peer groups with respect to students' predicted performance ("peer quality"). The results are as follows: First, students' outcomes are positively influenced by their peers' quality. Second, a simulation analysis shows that a policy maker who cares about average achievement should compose groups so that average peer quality balances across groups. Chapter 2 examines gender peer effects in the same context. The analysis shows that while female students seem to benefit from higher shares of females in their peer groups, no clear policy rule for gender group composition can be established.

Chapter 3 (co-authored with Darjusch Tafreschi and Sharon Pfister) examines the effects of course repetition in higher education. Students at the University of St. Gallen who do not reach a certain performance threshold have to repeat the full first year or to drop out otherwise. We compare individuals to both sides of this

threshold, but close to the threshold, to determine the effect of repetition. Repetition of a full year positively and persistently affects subsequent grades.

The last two chapters investigate labor supply decisions. Chapter 4 (co-authored with Christina Felfe and Michael Lechner) studies the impact of the availability of after-school care for schoolchildren on parents' employment and work hours in Switzerland. The analysis exploits variation in childcare density at the municipality level that is generated by differences in cantonal laws. We restrict the analysis to small regions at cantonal borders and compare only municipalities that are similar in observable characteristics. Higher childcare density results in higher shares of full-time employment for mothers of schoolchildren, which crowds out full-time employment for fathers of schoolchildren.

The final chapter, Chapter 5 (co-authored with Monika Bütler, Eva Deuchert, Michael Lechner, and Stefan Staubli), studies financial work incentives for disability insurance recipients in a randomized field experiment in Switzerland. Despite a substantial payment offered, the program is ineffective in inducing disability insurance recipients to work more and to rely less on benefits.

Zusammenfassung

Diese Dissertation besteht aus fünf Aufsätzen aus den Bereichen der Arbeitsmarkt- und Bildungsökonomie. Das Ziel der Dissertation ist es, ein Verstädnis für die Faktoren zu entwickeln, die individuelle Bildungsziele und individuelles Arbeitsangebot determinieren. Die Hauptmotivation dieser Arbeit liegt darin, dass Politikmassnahmen auf verschiedenen institutionellen Ebenen (z.B. auf der Ebene der Universität oder der Regierung) diese Faktoren zu einem gewissen Grad beeinflussen können.

Die ersten beiden Kapitel beschäftigen sich mit der Frage, welche Rolle Peer-Gruppen für post-sekundäre Bildungsergebnisse spielen. Für viele Studienanfänger stellen Freunde und Studienkollegen eine wichtige Hilfe und Informationsquelle dar. Um den Effekt der Zusammensetzung von Peer-Gruppen auf akademische Leistungen zu untersuchen, nutze ich Orientierungsgruppen an der Universität St. Gallen, die mit Hilfe eines Zufallsmechanismus zusammengestellt wurden. Kapitel 1 untersucht den Effekt der Gruppenzusammensetzung im Hinblick auf die vorausgesagten Leistungen der Mitstudierenden ("Peer-Qualität"). Die Analyse ergibt Folgendes: Erstens beeinflusst die Peer-Qualität die Leistung der Studierenden. Zweitens ergibt sich aus einer Simulationsanalyse, dass ein Entscheidungsträger, der den Leistungsdurchschnitt positiv beeinflussen möchte, die Gruppen so zusammensetzten sollte, dass die Peer-Qualität über die Gruppen hinweg ähnlich ist. Kapitel 2 untersucht die Effekte der Geschlechterzusammensetzung dieser Gruppen im gleichen Kontext. Die Analyse zeigt, dass Frauen von höheren Frauenanteilen leicht profitieren. Jedoch kann aus dieser Beobachtung keine klare Regel für die Zusammensetzung der Gruppen abgeleitet werden.

Kapitel 3 (mit Darjusch Tafreschi und Sharon Pfister) analysiert, wie sich Kurswiederholungen an der Universität auswirken. Studierende der Universität St.

Gallen, die eine bestimmte Leistungsschwelle nicht erreichen, müssen das erste Studienjahr wiederholen oder andernfalls ihr Studium beenden. Um den Effekt des Wiederholens zu analysieren, vergleichen wir Studierende, die sich diesseits und jenseits dieser Leistungsschwelle befinden, aber nah an der Leistungsschwelle sind. Das Wiederholen beeinflusst die Noten in den folgenden Studiensemestern positiv und nachhaltig.

Die letzten beiden Kapitel untersuchen Arbeitsangebotsentscheidungen. Kapitel 4 (mit Christina Felfe und Michael Lechner) analysiert den Einfluss der Verfügbarkeit von Mittags- und Nachmittagsbetreuung für Schulkinder auf die Erwerbstätigkeit der Mütter und Väter dieser Kinder in der Schweiz. Die Analyse nutzt Unterschiede in der Intensität des Betreuungsangebotes auf der Gemeindeebene, die wiederum durch Unterschiede in der Gesetzgebung zwischen den Kantonen hervorgerufen werden. Wir beschränken uns auf kleine Regionen in der Nähe der Kantonsgrenzen und vergleichen nur diejenigen Gemeinden, die sich im Hinblick auf beobachtbare Charakteristika ähneln. Eine höhere Betreuungsdichte führt dazu, dass ein höherer Anteil an Müttern Vollzeit erwerbstätig ist. Gleichzeitig reduziert sich der Anteil der Vollzeitbeschäftigten unter den Vätern.

Das letzte Kapitel, Kapitel 5 (mit Monika Bütler, Eva Deuchert, Michael Lechner und Stefan Staubli), untersucht finanzielle Arbeitsanreize für Personen, die Leistungen der Invalidenversicherung (Arbeitsunfähigkeitsversicherung) in der Schweiz beziehen, mit Hilfe eines Feldexperiments. Obwohl den Versicherten eine substanzielle Summe angeboten wurde, können diese nicht dazu veranlasst werden, mehr zu arbeiten und sich weniger auf Versicherungsleistungen zu stützen.

1. Social Planning with Spillovers: The Persistent Effects of Short-Term Peer Groups

Petra Thiemann

Abstract

This paper studies peer and reallocation effects of a one-week intervention in higher education. The analysis is based on a peer quality score. This measure integrates multiple exogenous peer characteristics into a single score, and therefore simplifies the analysis of peer and reallocation effects when various peer characteristics are potentially important. I use a dataset of six cohorts of freshmen (2003–2004 and 2006–2009) at the University of St. Gallen to study the impact of social ties that are formed during the first week at university on the probability of passing the freshmen year. The analysis exploits randomization of students into freshmen groups to identify peer and reallocation effects. The results are as follows: The paper finds significantly positive effects of an increase in peer quality on academic performance for students in the bottom quartile of the distribution of peer quality. The analysis also suggests detrimental, but only weakly significant, effects on average outcomes from an increase in segregation according to peer quality, compared to the status quo allocation. Furthermore, segregation seems to aggravate the gender gap in higher education outcomes.

JEL codes: I21, I23, J24

Keywords: peer effects, segregation, higher education

1.1 Introduction

When students enter college or university, they face many decisions that are crucial for their future career: which subjects to take, how, where, and when to study, how to spend their free-time, and whether to stay or to drop out, for example. Peer or reference groups influence these decisions and therefore students' outcomes in important ways (see Epple and Romano (2011) and Sacerdote (2011) for reviews). At the same time, the initial phase of entering university seems decisive for peer group formation. In this phase, even minor and sometimes random events like the seating order in a lecture can determine peer group formation, as shown in the social psychology literature (Back et al., 2008). Taken together, these two observations trigger the questions: Should university administrators care about and potentially intervene in the early peer group formation process in higher education? And if yes, which types of peer group assignments should administrators try to promote?

In order to characterize different types of potential assignments, or "reallocations" with respect to the status quo assignment, the distinction between integration versus segregation has become important both in the education policy debate and in the related economics literature (Graham et al., 2010, Graham, 2011). Integrating allocations mix individuals such that groups or classrooms are as homogeneous as possible in terms of individual characteristics, whereas segregating allocations do the opposite.[1] The prime example from the education literature, which is also the most extensively studied, is the effect of racial segregation in schools on students' achievement (Angrist and Lang, 2004, Card and Rothstein, 2007). Another example of segregation is tracking, i.e. sorting of students according to their performance into high versus low ability groups (Kremer et al., 2011). While these questions feature more prominently in the primary and secondary education literature, little systematic evidence about segregating versus integrating peer groups in higher education exists, mainly for two reasons. First, direct evidence on reallocation effects in higher education is almost absent from the literature, with the exception of Carrell et al.

[1] Likewise, marginal reallocations shift the allocation slightly toward more or less segregation, as outlined by Graham et al. (2010).

(2013) as well as Bhattacharya (2009).[2] Second, existing studies on peer effects in higher education have not systematically made use of the segregation-integration distinction to characterize beneficial policies.[3]

This paper complements the literature by studying peer effects and the effects of segregation in higher education. I focus on passing the first year at university as the outcome variable. Based on a binary choice model (logit), I implement a simulation-based approach to the study of segregation effects suited for experimental or quasi-experimental data. I use a unique dataset from the University of St. Gallen to infer peer effects and the effects of segregation at the onset of university education. Students spend the first full week of their first year (freshmen week) in groups of on average 15 students. Peer groups under the status quo can be considered random in terms of observable and unobservable characteristics, conditional on gender and admission rule (whether the student was required to take an entrance exam or not, coded as a dummy variable). Strong gender differences in academic performance - the probability of passing is on average by 6 percentage points lower for female students, compared to male students - motivate a particular focus on heterogeneous effects by gender, and on the effects of segregation on the gender gap in academic performance.

The decisive feature of the binary choice model used for the analysis is its way of combining multiple peer characteristics into a single score. Following Pinto (2011), the model relates the outcome to both individual student quality and a peer quality score. Student quality is defined as a weighted average of student characteristics, and peer quality is the average of student quality over all peers. Determining the weights on these characteristics in turn is data driven, i.e. determined when the model is estimated.[4]

[2] Many papers, e.g. Sacerdote (2001) and Lyle (2009), recognize the existence of beneficial reallocations. Yet, the authors do not directly derive reallocation effects.

[3] This literature also relates to the literature on optimal treatment allocation, see for example Manski (2004), Behncke et al. (2009).

[4] Pinto uses dimension reduction in order to implement a semi-parametric approach to the study of peer effects.

Dimension reduction simplifies the analysis of both peer and reallocation effects. First, dimension reduction allows for a sufficiently flexible modeling of peer effects, which allows, for example, for decreasing returns to peer quality. Second, reallocation effects are difficult to compute when multiple peer characteristics are relevant.[5] In most applications however, many different factors seem potentially important. One pragmatic solution to this problem is to pick one important variable, e.g. a variable that seems to generate large peer effects (Carrell et al., 2013). The drawback of this approach however is twofold: On the one hand, gains might be higher when considering multiple characteristics at once (Bhattacharya and Dupas, 2012, Bhattacharya, 2009). On the other hand, choosing the most important peer variable is not straightforward. For example, trade-offs between size and precision of the average marginal effect might become important.

Based on this model, the analysis proceeds in two steps. First, I define the average partial effect (APE) of peer quality as the expected effect of marginally increasing peer quality on the probability of passing the first year for an individual randomly drawn from the population. I compute the effect both for the whole sample and for different quartiles of the peer quality distribution separately[6] in order to account for non-linearities in the effect. Inference on peer effects relies on two alternative methods, both bootstrap and randomization inference. Second, I propose a simulation method to derive and evaluate a set of hypothetical reallocations. I characterize each reallocation based on its degree of segregation in terms of peer quality.[7] Inference on reallocation effects relies on a bootstrap method.

The results are as follows: First, the paper finds significantly positive effects of an increase in peer quality on academic performance for individuals below the median of the distribution of peer quality. For this group, an increase in peer quality by one standard deviation increases the average probability of passing the first year

[5] See Graham et al. (2010) as well as Bhattacharya (2009).

[6] The effect for different sub-samples, defined according to individuals' position in the peer quality distribution, can also be characterized as a local average response (LAR).

[7] Segregation is defined in terms of within-group-variance: The smaller this expression, the more segregated is the allocation.

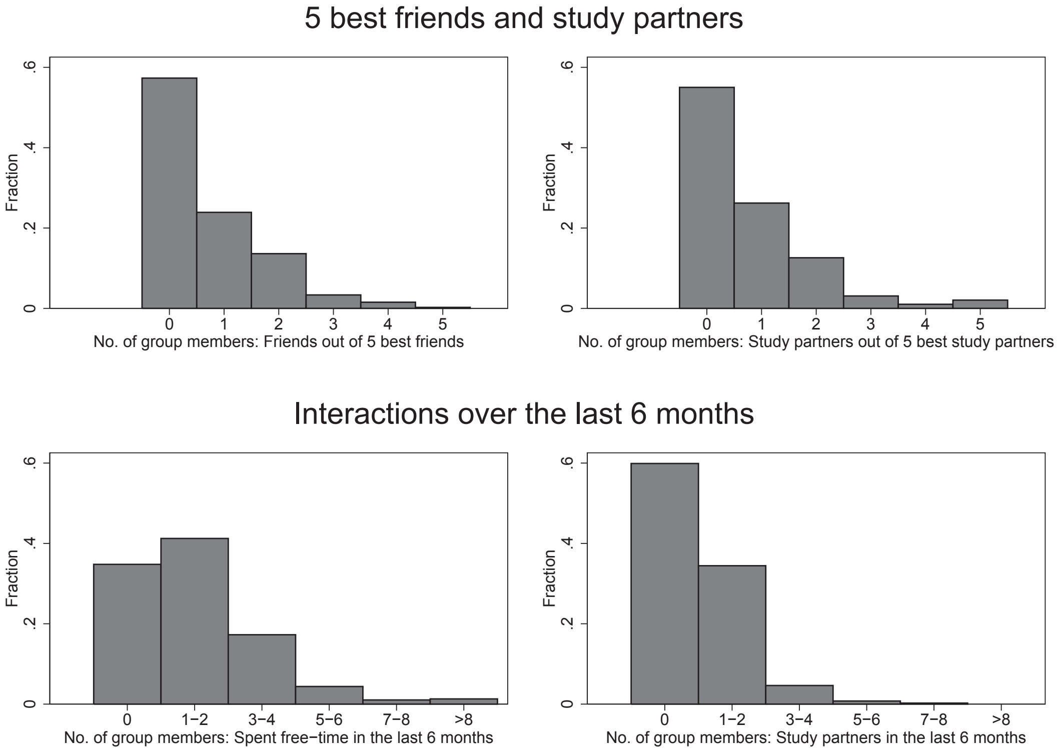
by 2.2 percentage points, which is sizable given an average passing rate of 66%. Second, splitting the sample according to gender, this result remains significant only for male students. Third, an increase in segregation decreases average student outcomes by up to 3.4 percentage points. This effect is only weakly significant. Fourth, an increase in segregation increases the gender gap in student outcomes by up to 5.3 percentage points. This is almost the size of the original gender gap (6 percentage points). The reallocation results, however, have to be qualified, as part of the simulated groups lie outside the support of peer quality in the status quo allocation. Overall, none of the simulated reallocations performs appreciably better than the status quo allocation, which is a highly integrated allocation, both in terms of maximizing average outcomes, and in terms of equalizing gender outcomes.

These results yield answers to the above questions. First, university administrators should care about peer group formation, especially in the starting phase of university education. Even relatively short interventions like an introductory week seem to influence subsequent outcomes through peer group assignment. Second, classifying reallocations in terms of their degree of segregation seems a useful approach to the characterization of beneficial policies. In the setting studied in this paper, more integrated allocations clearly outperform more segregated allocations.

The channels for these effects however remain the subject of further study. To get a first impression on the relationship between the intervention under study and the formation of friendships and study partnerships, I carried out an online-survey among all currently enrolled Bachelor students in Spring 2012.[8] For most of the students, the survey took place 1 to 3 years after their freshmen week. Therefore, results from this survey, although non-representative, provide insight into the long-run impact of the intervention on social interactions. Indeed, contacts from the freshmen week seem to last beyond the first semester, and even beyond the first year (see Figure 1.1). Furthermore, group members are over-represented among friends and study partners (see Section 1.3). The findings from the survey are therefore consistent with a friendship formation model where the probability of friendship formation between students from a common group is higher than the probability

[8] The survey was administered to 2,124 students and had a response rate of 18%.

Figure 1.1: Survey results on formation of friendships and study partnerships



Results from an online-survey, which was administered to all currently enrolled Bachelor students in Spring 2012. The survey had a response rate of 18%, with 388 students answering the above questions. For most students, the survey took place 1 - 3 years after their freshmen week. For detailed information on the survey questions, see Section 1.B.

of friendship formation between students from different groups. Thus, a randomly drawn student from a group with high peer quality may on average have higher quality friends, compared to a randomly drawn student from a group with low peer quality. Yet, the exact channel trough which peer effects operate remains a subject for further research.

The paper proceeds as follows: The following section gives an overview over the related literature, together with a preview on the contributions made by this paper. Section 1.3 explains the institutional background, followed by a descriptive overview of the data in Section 1.4. Section 1.5 introduces the model as well as measures of peer and reallocation effects. Section 1.6 presents the results, followed by a discussion of the results in light of the existing literature in Section 1.7. Section 1.8 concludes.

1.2 Related literature and contribution of this paper

This study builds upon a series of papers that study peer effects in higher education, and complements the existing literature in several dimensions, both with respect to the research questions, and with respect to the methodology used. First, this paper studies a novel setting, i.e. an introductory week, which is a relatively short intervention before undergraduate education starts. Settings studied before include: roommates (Sacerdote, 2001, Zimmerman, 2003, Stinebrickner and Stinebrickner, 2006, Kremer and Levy, 2008), dormitories (Foster, 2006), squadrons in a military academy (Lyle, 2007, 2009), cohorts at an air force academy (Carrell et al., 2009, 2013), members of the same workgroup (Oosterbeek and van Ewijk, 2014), and students attending the same classes or lectures (De Giorgi et al., 2010).[9] Overall, effects of roommates are relatively small, compared to the other settings that have been studied. According to Stinebrickner and Stinebrickner (2006), interactions between roommates arise out of necessity, and therefore roommate effects might not be as relevant as the effects of friends or study partners. The focus on alternative settings

[9] All of these studies focus on exogenous peer effects, with the exception of the paper by De Giorgi et al. (2010), that studies endogenous peer effects.

and larger peer groups therefore brought new insights and stronger results (Carrell et al., 2009). In terms of peer group size, the setup studied here is close to the paper by Carrell et al. (2009). While the authors study squadrons of on average 30 individuals, this paper examines groups of 15 individuals. Yet, the setting studied here is sufficiently distinct to merit further investigation. While students spend their whole first year at the air force academy entirely with members of their squadron, the orientation week facilitates initial contacts between students within groups, without preventing students from making contacts across groups during their first year.

Second, this paper studies an outcome that is relevant to lifetime earnings, i.e. the probability of passing the first year. Thus, it complements the literature which has focused primarily on test scores. Other outcomes studied so far for example include choice outcomes such as major choice (Lyle, 2007), joining a fraternity (Sacerdote, 2001), or drinking behavior (Kremer and Levy, 2008). While all these factors might contribute to labor market success, the outcome under study seems directly important for lifetime earnings. Repeating a year at the University of St. Gallen in case of failing the first year, and therefore delaying labor market entry by one year, corresponds to foregone earnings of 80,000 Swiss Francs (approximately 88,000 USD), which is the average entry salary of a newly graduated Bachelor student from the University of St. Gallen (Egger and Dyllick, 2010). To my knowledge, only Lyle (2007) and Oosterbeek and van Ewijk (2014) study similar outcomes, i.e. the decision to remain in the army for more than 6 years and the decision to drop out of university, respectively. According to Kremer and Levy (2008), choice variables seem more responsive to peer effects than pure performance outcomes. The probability of passing the first year includes a strong choice component. Passing the first year depends to a large extent on the decision not to drop out (see Section 1.4).

Third, this paper is one of the few studies investigating effect heterogeneity with respect to gender, as well as possibilities to close the gender gap by allocating individuals to peer groups. Oosterbeek and van Ewijk (2014) exploit randomization of the share of female students in tutorials, and find gender heterogeneity. Only boys' math performance decreases in the share of girls, but girls' performance remains unaffected. On the contrary, Arcidiacono and Nicholson (2005) study peer effects within classes of cohorts in medical schools, and find peer effects only for female students. [10] Due to these mixed results, adding to the literature on effect heterogeneity by gender seems worthwhile. In addition, this study is the first to explicitly investigate how the gender gap in academic outcomes responds to peer group allocations. This topic is especially important in light of the debate on the gender gap in labor market outcomes. As discussed by Bertrand et al. (2010) for the case of MBA students, a substantial part of the gender earnings gap may come from differences in course choices and academic performance at university. Therefore, diminishing the gender gap in academic outcomes may directly contribute to a smaller gender gap in labor market outcomes.

Fourth, this study builds upon Pinto's idea of using an index variable that combines various peer characteristics into a single score (Pinto, 2011). Previous studies have been interested in the effect of multiple peer characteristics as well. They examine these variables either in separate regressions, or jointly in a single regression. While these approaches are intuitive, each of them brings about disadvantages. On the one hand, when examining each characteristic separately, peer effects of either variable might appear small, which can be misleading when their joint impact is strong. On the other hand, coefficients from multivariate analyses of several characteristics in one regression might be difficult to interpret, especially when these are strongly correlated. Marginally changing one characteristic while holding the other characteristics of the peer group constant is then hardly feasible. Introducing a score addresses both concerns. First, the score might have a stronger impact, and contains more information, compared to each single characteristic alone, and second, average partial effects allow for a clear interpretation (see Section 1.5).

Fifth, this study also focuses on the effects of reallocations. Some existing studies already go beyond the question whether peer effects exist. They examine whether rearranging peer groups would enhance aggregate welfare, e.g. in terms of academic performance (Sacerdote, 2001, Carrell et al., 2009, Lyle, 2009, Carrell et al., 2013). The roommate studies deal only with pairs of peers. In these contexts, conclusions

[10] Two roommate studies investigating gender heterogeneity are Stinebrickner and Stinebrickner (2006) and Zimmerman (2003).

about beneficial reallocations of peers can be drawn relatively easily on the basis of effect heterogeneity analyses (Sacerdote, 2001, Lyle, 2009).[11] Moreover, Bhattacharya (2009) presents a rigorous framework for analyzing reallocations in roommate settings in a linear programming framework. The case of groups that are larger than two is more difficult to tackle and requires further assumptions, as outlined by Graham et al. (2010). Building upon these new methodological approaches, Carrell et al. (2013) not only determine an optimal peer allocation in larger groups, but also implement this allocation in a controlled setting. Yet, the previously computed optimal assignment does not lead to better aggregate outcomes. Consequently, the issue of optimal peer groups remains a field that requires further examination. The method implemented in this paper relies on simulation of a set of reallocations that differ in terms of their segregation in terms of observable characteristics (see Section 1.5).

1.3 The freshmen week at the University of St. Gallen

The University of St. Gallen in Switzerland offers undergraduate studies in Business Administration, Economics, International Affairs, Law and Economics, as well as Legal Studies. Undergraduate degrees take a minimum of three years to complete. The first year serves as a selection device and orientation period. Almost all first-year students complete the same set of classes, with minor exceptions.[12] Academic performance by the end of the first year determines whether students are admitted to the second year. On average, 66% of students pass the first year in their first attempt (see Table 1.1). All other students either drop out beforehand or fail the exams.[13] After the first year, students choose their major.

[11] Yet, none of these studies presents rigourous identification, estimation, and inference results on optimal peer allocations.

[12] Exceptions include: Students with foreign mother tongue who choose to complete all first year courses within two years (extended track), and students majoring in Legal Studies. For the latter group, 2 first-year courses differ. Both groups combined account for 13% of freshmen, see Table 1.1.

[13] Students who fail can repeat all first year courses in order to be admitted to the second year.

Undergraduate studies start with a mandatory freshmen week. This week familiarizes students with university infrastructure (e.g. library and online tools), facilitates contacts between students, and introduces them to team work, which makes up an important part of the studies at the University of St. Gallen. Students are sorted into teams at the beginning of the week. Each team consists of, on average, 15 individuals, with between 57 and 60 teams per cohort (see Table 1.A.2). Group sizes and the number of groups can vary between cohorts. Moreover, variations of group sizes within cohorts emerge from organizational constraints (in particular, room size). During the week, students spend approximately 60 hours in their groups, with about 75% of the time dedicated explicitly to team activities. A case study competition between groups forms the core team activity.

The assignment mechanism to freshmen groups ensures random assignment of peers in terms of observable and unobservable characteristics, conditional on gender and the admission protocol that applies to the respective student.[14] More precisely, the university implements the following three-step mechanism. First, students are divided into four strata according to gender (male vs. female) and entry requirement (exam vs. no exam). Second, each stratum is sorted according to students' surnames. Third, student 1 of stratum 1 is placed into the first group, student 2 of stratum 1 into the second group, and so forth, until the stratum is empty. Then, the process starts again: Student 1 of stratum 2 is placed into the first group, student 2 of stratum 2 into the second group, and so forth. This mechanism is repeated for all 4 strata. In this way students with identical or similarly starting surnames most likely end up in different groups. Later on, an administrator might re-distribute peers across groups in order to match group sizes to available room sizes. This redistribution occurs unsystematically.

The freshmen week facilitates the formation of friendships and study partnerships among undergraduates, as a survey among a cross section of Bachelor students

[14] All individuals with a non-Swiss high school diploma and non-Swiss nationality face an entrance exam as additional barrier to entry.

(second year and above) in 2012 shows.[15] In the sample of 389 survey respondents, the probability of a student having at least one freshmen team member among his 5 best friends amounts to 0.45 (with a standard deviation of 0.025). Likewise, the probability of a student having at least one study partner among his 5 most frequent study partners amounts to 0.42 (with a standard deviation of 0.025). These results indicate the presence of a positive impact on social ties: The probability of these events occurring at random in a cohort of 1,000 students and with a group size of 15 team members amounts to only 0.072.[16] If students select into survey participation based on their friendship with freshmen week peers, the survey results might however overstate the impact on social tie formation. Furthermore, only a small number of freshmen week contacts remains important. The probability to be friends or to study with more than two team members during the last 6 months amounts to only 25% or 6%, respectively. To sum up, freshmen groups facilitate social tie formation, but these social ties stabilize only among very small subgroups, or even only among pairs of peers.

1.4 Data and descriptive statistics

1.4.1 Sample

The dataset consists of administrative records for 5,024 freshmen starting their undergraduate studies in 2003 - 2004 and 2006 - 2009.[17] Background (pre-treatment) characteristics as well as outcomes are computed from enrollment and grade records. Freshmen group assignment can be matched to these records based on a student identifier. Only a few first-year students had to be deleted from the sample: Students who could not be identified in the freshmen group file, partly because they did not

[15] We carried out the survey for the purpose of this research project in May 2012. The online survey was administered to all 2,124 currently enrolled Bachelor students (second year and above) and had a response rate of 18%.

[16] P[At least 1 freshmen group friend] $= 1 -$ P[No freshmen group friend] $= 1 - \frac{999-15}{999} * \frac{998-15}{998} * \frac{997-15}{997} * \frac{996-15}{996} * \frac{995-15}{995} \approx 0.072$.

[17] Freshmen group data are unavailable for the year 2005.

participate, as well as students who participated in special (self-)selected freshmen groups.[18] Consisting of 97% of freshmen, the sample is largely representative for the freshmen student body (see Table 1.A.1).

1.4.2 Background characteristics

Available pre-treatment characteristics, measured before entry, come from enrollment records. Available characteristics include: age, gender, admission protocol, mother tongue (German vs. non-German), nationality, and country of high school degree. Furthermore, the characteristics include variables on two special tracks: First, individuals with non-German mother tongue can choose to complete all first-year courses within two years. Second, individuals in the legal studies track take two special classes. All other students complete the exact same set of classes.

These characteristics depict the homogeneity of the student body (see Table 1.1), mostly in three aspects. First, the vast majority of students are male (69%). Second, only a minority of students come from foreign countries (24%) or have a foreign high school degree (23%). Individuals fulfilling both criteria at once have to pass an entrance exam to be admitted, due to Swiss legislation (18%).[19] Third, although Switzerland has 4 different official languages (German, French, Italian, Rumantsch), only 11% of students speak a non-German mother tongue, mainly because all first-year courses are taught exclusively in German.

These exogenous characteristics are strongly interrelated, as shown in a correlogram (Table 1.A.4). Some of the correlations are driven by institutional entry requirements (e.g. the positive correlations between admission protocol, foreign nationality, foreign high school degree, and extended track). Foreigners also less often choose the legal studies track, as this major focuses on the Swiss legal system. Gender is also strongly correlated with other characteristics: Male students are more likely to have a foreign passport or a foreign high school degree. Moreover, male

[18] Students who had to serve in the army at the time of the freshmen week formed a special group, and up to 3 groups per semester are groups with special tasks ("media groups") for which students could sign up beforehand.

[19] This legislation restricts access for foreign students to at most 25%.

Table 1.1: Descriptive statistics for the estimation sample

Variable	Mean	Std.	Min.	Max.
Pre-treatment characteristics				
Male	0.69	-	0	1
Admission protocol$^{(1)}$	0.18	-	0	1
Age (years)	20	1.93	16	48
Non-German mother tongue	0.11	-	0	1
Legal studies track	0.07	-	0	1
Non-Swiss nationality	0.24	-	0	1
Non-Swiss high school degree	0.23	-	0	1
Extended track	0.06	-	0	1
Student (own) quality	0.03	0.55	-2.14	1.07
Peer variables and treatment				
Groupsize	15.20	3.08	7	22
Share of peers: male	0.69	0.08	0.43	1
Share of peers: admission protocol	0.18	0.07	0	0.45
Share of peers: 20 years and older	0.30	0.13	0	0.78
Share of peers: non-German mother tongue	0.11	0.09	0	0.44
Share of peers: legal studies track	0.07	0.08	0	0.42
Peer quality	0.03	0.13	-0.44	0.45
First year outcomes				
Voluntary dropout during 1st semester	0.07	-	0	1
Failed during 1st semester	0.11	-	0	1
Voluntary dropout during 2nd semester	0.07	-	0	1
Failed during 2nd semester	0.09	-	0	1
First year passed successfully	0.66	-	0	1

Descriptive statistics for the estimation sample (5,024), based on administrative student records for cohorts 2003 - 2004 and 2006 - 2009. Peer quality as well as student (own) quality are computed using the Model outlined in Section 1.5.1. First year outcomes presented here are mutually exclusive. The outcome "First year passed successfully" will be used as outcome variable throughout the further analysis.

(1) Admission protocol = 1 if the student has to pass an entrance test to be admitted.

students are older as most of them serve in the army before starting their degrees. Almost counter-intuitive at first sight seems the correlation between mother tongue and nationality: Students from foreign countries are less likely to speak a non-German mother tongue, as they come primarily from German speaking neighboring countries (Germany, Austria).

1.4.3 Outcome

As the first year serves as a (self-)selection period, I define the outcome as whether individuals pass the first year in their first attempt.[20] This outcome serves as indicator for academic success, and is defined for all individuals in the sample. After successful completion of the first year, most students also successfully complete their Bachelor degree.[21] Grades of first-year courses do not count towards the grade of the Bachelor degree. Several reasons for not completing the first year successfully in a first attempt exist: Individuals drop out (7% per semester, see Table 1.1), or fail the first year (11% fail during the first semester, and 9% fail during the second semester, respectively, see Table 1.1). Individuals who do not pass the first year can decide to repeat the full first year, or to switch university.[22] To sum up, while the outcome under examination - finishing the first year in the first attempt - predicts academic success at the University of St. Gallen well, it provides only incomplete information about academic performance in general.

Table 1.A.3 shows strongly significant gender differences in the outcome under examination. Only 61% of females pass the first year in their first attempt, compared to 68% of males. Females not only fail more frequently, but also drop out voluntarily at a higher rate. Gender differences in background characteristics parallel gender differences in the outcomes, which might partly explain the gender gap. Further

[20] For individuals in the extended track, the outcome corresponds to successful completion of first-year courses within two years.

[21] E.g., in the cohort of 2004 (2006), 96% (93%) of students who pass the first year in their first attempt obtain their Bachelor degree within at most 5 years. This information is unfortunately not available for all cohorts due to data censoring.

[22] For example, in the cohorts 2004 and 2006, 27% of students who fail their first attempt still manage to finish their Bachelor degree within 5 years at the University of St. Gallen.

investigations into the drivers of the gender gap however go beyond the scope of this paper.[23]

1.4.4 Peer characteristics

All background characteristics described above are potential drivers of exogenous peer effects, however I include only a subset of available peer characteristics into the model. The variables I include are dummy variables for: gender, admission protocol (= 1 if entrance exam required), age (20 years and older vs. younger than 20 years), German mother tongue, and major (= 1 if legal studies track). Due to high correlation with these variables, I exclude all other available background characteristics from the estimation. For all five relevant peer variables, I create leave-own-out group means for each student.

One way to test for exogeneity of these treatment variables with respect to individual background characteristics is to regress individual characteristics on peer characteristics. Under random assignment, peer characteristics cannot predict individual characteristics. As group assignment is only conditionally random, I regress individual characteristics (age, non-German mother tongue) on peer characteristics within strata (see Table 1.A.5). In seven of eight regressions, peer characteristics are jointly insignificant both at the 5% and at the 10%-level. In one regression, peer characteristics are significant at the 10%-level. These results support the assumption of independence of treatment and observable individual characteristics.[24]

The model presented in Section 1.5.1 aims at reducing the dimension of the treatment variable from a vector of treatment characteristics to a scalar treatment variable (here: from dimension 5 to dimension 1). This dimension reduction is data driven, i.e. happens within the model. The new variable is a weighted average of the 5 treatment variables. I refer to the aggregated treatment as "peer quality" (see Section 1.5 for further explanations).

[23] For a comprehensive survey on the drivers of gender differences in labor market related outcomes, see Bertrand (2011).

[24] I do not regress on the variable "legal studies track" due to small variance of this variable within some of the strata.

1.5 Measuring peer and reallocation effects

1.5.1 Measuring peer effects

The model

The model relates the outcome Y_{ig} of student i in group g (whether a student passes the first year in his first attempt) to mean peer characteristics and individual characteristics in the following way:

$$\mathrm{P}[Y_{ig} = 1 | X_{ig}, X_{-ig}] = F_{mc}(X'_{ig}\beta, \overline{X}'_{-ig}\beta), \tag{1.1}$$

where X_{ig} is a vector of individual characteristics, $\overline{X}_{-ig}$ is a vector of mean peer characteristics of student i in group g, where the mean is computed over all students in the same group, excluding the student himself, and F_{mc} is a function (to be identified) that potentially differs across gender (index $m \in \{0, 1\}$ with $m = 1$ if male and $m = 0$ if female) as well as cohort (index c). In this model, $\overline{X}'_{-ig}\beta$ – a scalar for each individual – contains all information necessary to determine peer effects, as proposed by Pinto (2011). $\overline{X}'_{-ig}\beta$ is a weighted average of peer characteristics. The weights used in order to integrate peer characteristics into a single score correspond to the coefficients on the respective individual characteristics. Henceforth, I refer to the score $\overline{X}'_{-ig}\beta$ as "peer quality". Accordingly, I refer to $X'_{ig}\beta$ as "own quality" (or "student quality").

The aggregation scheme chosen to create the score comes from a simple intuition. First, the most important drivers of individual performance are also the most important drivers of peer effects. For example, individuals who have completed an entrance test might be highly motivated students, which might not only affect their performance, but might also spill over to their peers. This characteristic might be relatively more important than the choice of major, both for individual performance and for peers' performance. Second, not only the relative magnitudes but also the signs are identical for individual and peer characteristics. I.e. if the spillover effect is positive, every characteristics that makes a good student also makes a good peer (e.g. through knowledge spillovers). Likewise, if the spillover effect is negative, every characteristic that makes a good student also makes a bad peer (e.g. through discouragement).

This model builds upon the reduced form of the linear-in-means model first introduced and discussed by Manski (1993) (see, for example, Lyle (2007) for a derivation of the reduced form from the structural equation). But – given the setup discussed in Section 1.4 – why would peer effects at all be driven by group means? In particular, if students form persistent friendships or study partnerships with only a subset of the group, i.e. 1 or 2 students, why should we consider group means as the drivers of peer effects? Two main reasons speak in favor of the model. First, as most papers use mean peer characteristics as peer variables, regardless of the friendship formation process in the respective settings, this approach serves as a natural starting point and makes the results of this study comparable to other studies. Second, the model can be interpreted in terms of an "availability effect" as outlined by Carrell et al. (2013): If friendships and study partnerships were formed randomly within freshmen groups, students with high peer quality end up on average with higher quality friends and study partners, compared to students with low peer quality. Carrell et al. (2013) attribute more than two thirds of friends' characteristics to this availability effect. However, the authors also emphasize the importance of homophily, which seems to explain the remaining one third of friends' characteristics; students do not pick their friends and study partners randomly, but instead they tend toward making friends with similar attributes. Therefore, too many low ability students in one group render the effect of high ability peers negligible in their experiment, a mechanism that has not been caputed by their original model. In the model presented here, not considering homophily could also jeopardize the policy conclusions from the results: Groups with the very same average peer quality can have different within-group peer quality distributions, and therefore different potentials for finding a similar friend in terms of peer quality. While the model presented above still serves as a good starting point to measure the availability effect, this caveat has to be taken into account when deriving and interpreting assignment policies.

To tailor the model to the specific application in this paper, I include the following assumptions. First, the influence of individual characteristics and mean peer characteristics is additively separable inside an index. Additive separability is the standard assumption of the linear-in-means model, introduced by Manski (1993). Second, the influence of differences between cohorts also enter in an additively separable way. Third, the influence of peer quality on individual outcomes can potentially differ between males and females. Forth, as the outcome is binary, the model links the characteristics and the outcome through the logistic function. I therefore model the probability to pass in the following way:

$$\mathrm{P}[Y_{ig} = 1 | X_{ig}, X_{-ig}, D_{ig}] = \Lambda \left(X'_{ig}\beta + f_m(\overline{X}'_{-ig}\beta) + D'_{ig}\delta \right), \tag{1.2}$$

where D_{ig} is a vector of cohort dummies, $\Lambda(.)$ is the logistic function, and $f_m(.)$ is a potentially non-linear function, that differs by gender (to be specified below).

The following assumptions refer to the specification of $f_m(.)$. In order to approximate the potentially non-linear response of individual outcomes to peer quality (in addition to the non-linearity of the logit-model), the model introduces higher order terms for peer quality. Moreover, all coefficients on the peer variables can differ between male and female students, so that the model reads:

$$\mathrm{P}[Y_{ig} = 1 | X_{ig}, X_{-ig}, D_{ig}] = \Lambda \left(X'_{ig}\beta + \sum_{k=1}^{K} (\gamma_{1k} + \gamma_{2k}\mathrm{male}_i)(\overline{X}'_{-ig}\beta)^k + D'_{ig}\delta \right), \tag{1.3}$$

where male_i is a gender dummy. The specification used in this paper sets $k = 3$ in order to ensure sufficient flexibility and precision. The choice of $k = 3$ should be checked against alternative specifications. In principle, putting less ex-ante restrictions on the function $f_m(.)$, i.e. estimating the function nonparametrically, would be desirable. This is however not feasible in the current application due to data limitations, as most of the variation comes from between-group differences in peer quality.

Parameters of interest

In the following, I propose various parameters of interest. In defining these parameters, I follow the concepts of average partial effects (APE) (see, for example, Wooldridge (2005) and Wooldridge (2010)), of local average responses (LAR) (see, for example Altonji and Matzkin (2005)), and of the average structural function (ASF) (see, for example Blundell and Powell (2003)). First, the average partial effect of peer quality on individual outcomes can be defined as:

$$\begin{aligned}\theta &= E\left[\frac{\partial \mathrm{P}[Y_{ig}=1|X_{ig},X_{-ig},D_{ig}]}{\partial(\overline{X}'_{-ig}\beta)}\right] && (1.4)\\ &= E\left[\pi_i(1-\pi_i)\sum_{k=1}^{K}k(\gamma_{1k}+\gamma_{2k}\mathrm{male_i})(\overline{X}'_{-ig}\beta)^{(k-1)}\right], && (1.5)\end{aligned}$$

with $\pi_i = \mathrm{P}[Y_{ig} = 1|X_{ig}, X_{-ig}, D_{ig}]$. This quantity corresponds to the following thought experiment: If we were to randomly draw an individual from the population of students, by how much does the average outcome change in expectation if peer quality rises marginally? As discussed by Graham et al. (2010), this thought experiment does not correspond to an implementable policy under the assumption of a fixed student pool. An increase in peer quality for any individual corresponds to a decrease in peer quality for another individual if group sizes are fixed. Yet, the average partial effect, defined in this way, is interesting as it is informative about the overall sign and magnitude of the spillovers, which can already convey a first idea on whether policy makers should take spillovers into account.

The average partial effect can potentially differ with respect to gender. Therefore, I define the average partial effect for males and females, respectively, as

$$\begin{aligned}\theta_m &= E\left[\left.\frac{\partial \mathrm{P}[Y_{ig}=1|X_{ig},X_{-ig},D_{ig}]}{\partial(\overline{X}'_{-ig}\beta)}\right|\mathrm{male_i}=m\right] && (1.6)\\ &= E\left[\left.\pi_i(1-\pi_i)\sum_{k=1}^{K}k(\gamma_{1k}+\gamma_{2k}m)(\overline{X}'_{-ig}\beta)^{(k-1)}\right|\mathrm{male_i}=m\right], && (1.7)\end{aligned}$$

with $m = 1$ if the student is male, and $m = 0$ if the student is female. The corresponding thought experiment is analogous to the thought experiment for the previous parameter, but now we imagine a random draw only from the population of male or female students, respectively.

Second, potential presence of non-linearities in peer quality motivates breaking down the average partial effects further. For example, one might expect diminishing returns to average peer quality. Various studies emphasize the presence of such non-linearities (see, for example, Carrell et al. (2013)). One way to address these non-linearities is to split the sample into quartiles of the peer-quality-distribution and to look at the effect of peer quality on individual outcomes within each quartile. As the respective sub-populations are now defined by their values of peer quality, I refer to the parameters as a local average responses (LAR) according to Altonji and Matzkin (2005). Furthermore, I define these parameters for males and females separately. As an examle, I present here the expected effect for a randomly drawn individuals in quartile 1 (Q_1):[25]

$$\theta_{Q_1} = E\left[\frac{\partial \mathrm{P}[Y_{ig}=1|X_{ig}, X_{-ig}, D_{ig}]}{\partial(\overline{X}'_{-ig}\beta)}\middle|\overline{X}'_{-ig}\beta \in Q_1\right]. \tag{1.8}$$

.

Third, to fully understand non-linearities in the response to peer quality, we might be interested in the average structural function as defined by Blundell and Powell (2003). The value of ASF(q) denotes the value of expected outcomes under a counterfactual assignment. I.e., the value of ASF(q) can be interpreted as the expected outcome for a randomly drawn individual if this individual were assigned peer quality q.

$$\mathrm{ASF}(q) = E[\mathrm{P}[Y_{ig}=1|X_{ig}, (\overline{X}'_{-ig}\beta) = q, D_{ig}]] \tag{1.9}$$

As noted by Blundell and Powell (2003), if q can be manipulated directly, the ASF is sufficient to evaluate a policy: The optimal q is the one that leads to the highest expected outcome. Yet, in the peer effects setting studied here, the manipulation of

[25] Differences in the treatment effect between the quartiles can arise from two sources: First, from the shape of the response function to peer effects, and second, from differences between students from different quartiles with respect to individual characteristics. Due to random assignment, the first effect is supposedly the major driver of differences between quartiles in this application.

q is constrained by the distribution of peer quality in the student pool. Therefore, the ASF is indicative of the optimal allocation, but not sufficient to find an optimal policy. In particular, the ASF can be informative on which allocations are clearly dominated.[26]

Finally, I propose to estimate the derivative of the average structural function. This derivative captures the change in expected outcomes as a response to a marginal increase of q.

$$\mathrm{ASF}'(q) = E\left[\frac{\partial \mathrm{P}[Y_{ig} = 1 | X_{ig}, (\overline{X}'_{-ig}\beta) = q, D_{ig}]}{\partial q}\right]. \tag{1.10}$$

As Model 1.3 allows for effect heterogeneity with respect to gender, the ASF as well as its derivative are defined also for males and females separately.

Estimation and inference

I estimate the coefficient vector (β, γ, δ) using maximum likelihood estimation. Specifically, I maximize the natural logarithm of the likelihood function with respect to β, γ, and δ in order to find the corresponding coefficient vector:

$$(\hat{\beta}, \hat{\gamma}, \hat{\delta}) = \arg\max_{\beta,\gamma,\delta} \ln \mathcal{L}(\beta, \gamma, \delta | y_i, X_{ig}, D_{ig}), \tag{1.11}$$

where

$$\begin{aligned}\ln \mathcal{L}(\beta, \gamma, \delta | y_i, X_{ig}, D_{ig}) = &\sum_{i=1}^{N} y_i \ln\left[\Lambda\left(X'_{ig}\beta + \sum_{k=1}^{K}(\gamma_{1k} + \gamma_{2k}\mathrm{male}_i)(\overline{X}'_{-ig}\beta)^k + D'_{ig}\delta\right)\right] \\ &+ \sum_{i=1}^{N}(1 - y_i) \ln\left[1 - \Lambda\left(X'_{ig}\beta + \sum_{k=1}^{K}(\gamma_{1k} + \gamma_{2k}\mathrm{male}_i)(\overline{X}'_{-ig}\beta)^k + D'_{ig}\delta\right)\right].\end{aligned} \tag{1.12}$$

[26] For example, completely segregated allocations are clearly dominated when the ASF has an inverted u-shape.

I use the *Matlab* optimization procedure *fminunc* for estimation. To estimate the parameters described in 1.5.1, as well as the average structural function, I compute their sample analogues by plugging in the coefficients obtained from the estimation.[27]

Inference is based on two types of permutation methods: bootstrap inference and randomization inference (Fisher, 1971).[28] Both rely on creating reference distributions. Bootstrap inference is based on a bootstrap (resampling) procedure, which accounts for sampling uncertainty. I think of the sample of freshmen as a sub-sample coming from a (super-)population of potential students. Sampling from a larger population generates uncertainty with respect to the coefficient vector (β, γ, δ). The bootstrap procedure accounts for this type of uncertainty in the following way: First, I draw a number of bootstrap samples by sampling g_c groups with replacement for each cohort c (g_c denotes the number of groups in cohort c under the status quo allocation). Second, I obtain an estimate for the coefficient vector (β, γ, δ) for each bootstrap replication by estimating Model 1.3. Third, I compute the average partial effects as well as the average structural function within each bootstrap sample. I evaluate the bootstrap distribution as well as the average structural function, i.e. I compute their mean as well as 95% and 90% confidence intervals. This procedure allows to test various hypotheses, for example, the null hypothesis of zero peer effects, i.e. $H_0 : \hat{\theta} = 0$.

Randomization inference presents a second way to derive inference on peer effects, and accounts for the uncertainty generated by assigning individuals to their status quo freshmen groups. Again, I test whether peer quality has no influence on the outcome, i.e. $H_0 : \hat{\theta} = 0$. We know that the status quo assignment has come about as the result of a stratified randomization. But this status quo is just one potential assignment out of a large sets of counter-factual (or "placebo") assignments, which could have been achieved under the same stratified randomization scheme. The relationship between peer quality and the probability of passing under the status quo, which I denote here as $\hat{\theta}_{sq}$, could have come about completely by

[27] For the estimation of the average structural function, I ensure that q takes on only values within the support of $\overline{X'_{-ig}\beta}$ in the original sample.

[28] See also Ernst (2004) for an introduction, overview and comparison of these methods.

chance. To find the probability of detecting an effect as large or larger than $\hat{\theta}_{sq}$ just by chance (i.e. the p-value of the randomization test), randomization inference proceeds as follows. First, I define the "randomization distribution" as the distribution of $\hat{\theta}$ under all potentially possible counter-factual assignments coming from the same randomization scheme as the status quo assignment. The aim of randomization inference is to evaluate $\hat{\theta}$ versus this reference distribution. Second, in order to approximate the randomization distribution, I generate a random subset of all possible counter-factual assignments, keeping the strata proportions in each group fixed. I then compute the coefficient vector (β, γ, δ) and subsequently $\hat{\theta}$ for each counter-factual assignment, using Model 1.3. Third, I calculate the p-value of the randomization test of H_0. The p-value indicates the probability of finding an average partial effect of the size of $\hat{\theta}_{sq}$, or larger, if we were to draw a random $\hat{\theta}$ from the randomization distribution.[29]

1.5.2 Measuring reallocation effects

In order to derive reallocation effects, I construct a set of 56 hypothetical allocations that differ in terms of their within- and between-group variation in peer characteristics. Construction of these allocations proceeds sequentially, starting from a fully segregated allocation (allocation 0) and ending with a strongly integrated allocation (allocation 55). I never mix individuals across cohorts, so that allocations could indeed have been implemented in each year, and I preserve the distribution of group sizes within each cohort.

I construct this family of allocations as follows. Suppose that size of group 1 is g_1. A fully segregated allocation can be constructed by sorting individuals according to their predicted quality, and placing individuals $1, ..., g_1$ in group 1, individuals $(g_1 + 1), ..., g_1 + g_2$ in group 2, and so forth. Then, I construct a new allocation (allocation 1), which is slightly more segregated: I dissolve the first group (the group with the lowest average peer quality) and place all its individuals in different groups. All other individuals are again sorted according to their own quality and

[29] Performing a two-sided test, I assess the probability of the absolute value of $\hat{\theta}$ being larger than $\hat{\theta}_{sq}$.

filled into the remaining slots accordingly. I proceed with this mechanism until all initial groups from the most segregated allocation are resolved. This mechanism results in as many allocations as groups. As I do not mix cohorts across allocations, I end up with 56 allocations, as 56 is the minimum number of groups in a cohort.[30]

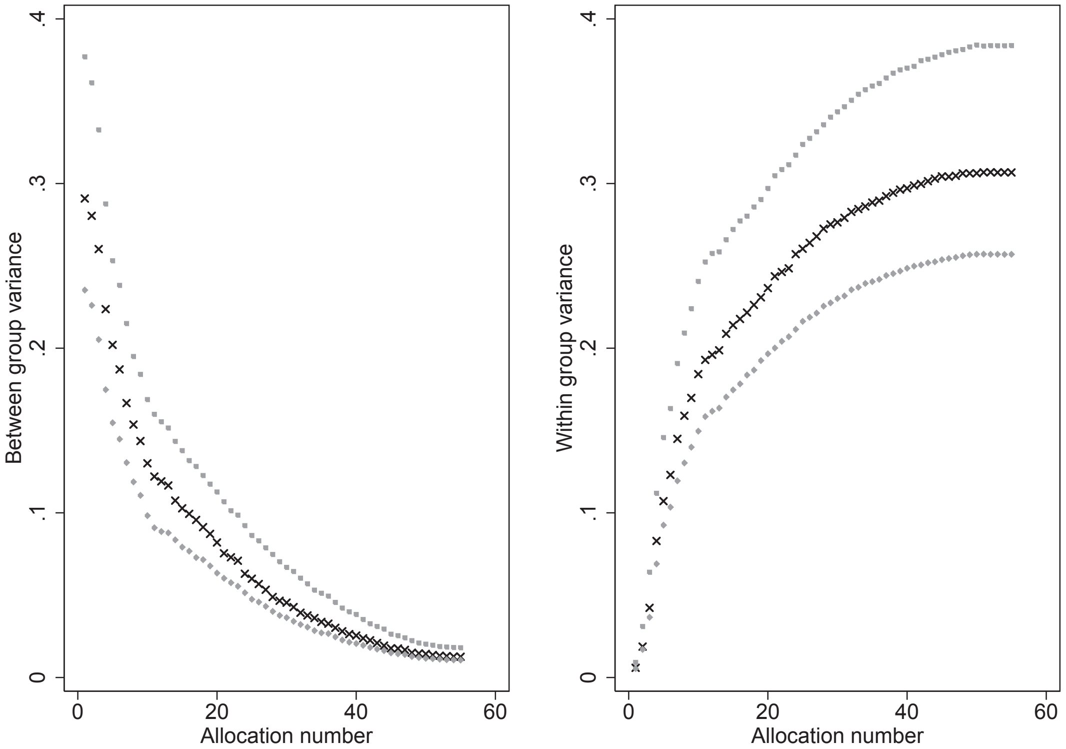
The chosen algorithm generates a family of allocations that differ in terms of their within- and between-group-variance in peer quality (Figure 1.2). This set of allocations is certainly not exhaustive. Moreover, within- and between variances are not sufficient to fully characterize an allocation in terms of observable characteristics. The goal of the analysis however is not to create, say, a unique mapping of within- or between-variances into average outcomes. Instead, the aim of the analysis is to understand the potential magnitude of reallocation effects in the given setting, by creating a family of sufficiently distinct allocations, and to understand part of the properties of beneficial allocations.

Reallocation gains (losses) can be expressed by comparing average outcomes under hypothetical allocations to average outcomes under the status quo allocation. In order to compute average outcomes for the 56 allocations described above, I use a plug-in procedure, drawing upon the estimated coefficient vector $(\hat{\beta}, \hat{\gamma}, \hat{\delta})$. I proceed in three steps. First, I predict peer quality for each hypothetical allocation, which results in a vector of 56 treatment variables $(\overline{X}'_{-ig}\hat{\beta}^{a0}, ..., \overline{X}'_{-ig}\hat{\beta}^{a55})$ for each individual. Second, I use the estimated coefficient vector γ, but plug in the hypothetical treatment variables $(\overline{X}'_{-ig}\hat{\beta}^{a0}, ..., \overline{X}'_{-ig}\hat{\beta}^{a55})$ to predict average outcomes under each new allocation. Third, in order to derive reallocation gains (losses), I compute the difference in average outcomes between each allocation and the status quo allocation. Using the same procedure, I also compute the gender gap under different allocations as well as their differences to the gender gap under the status quo.

The mechanism described above naturally generates a set of allocations which also includes extreme allocations. In particular, the first allocations contain (almost) perfectly segregated groups in terms of predicted peer quality, whereas the last allocations are almost perfectly integrated. Consequently, the values of peer quality in these groups might fall out of the support of peer quality in the status

[30] For further details on the assignment mechnism, see Table 1.C.14.

Figure 1.2: Between and within group variances for a set of hypothetical allocations



Left panel: Between-group variance in peer quality across 56 allocations. Right panel: Average within-group variance in peer quality across the same set of allocations. Allocations are computed using an algorithm creating first a fully segregated allocation with respect to peer quality (allocation 0), and then successively relaxing the degree of segregation across further allocations. See Section 1.C for details on the algorithm and on the computation of between- and within-group variances. The small, dotted lines refer to bootstrap confidence levels of within- and between-group variances (95% confidence interval) derived from the bootstrap procedure explained below. Based on 250 bootstrap replications and a sample of 5,024 observations.

quo allocation. Table 1.A.9 illustrates this point. Within the 56 allocations under examination, the first 28 allocations contain values of the treatment variable higher than the upper limit of the support of peer quality in the status quo allocation. The first 54 allocations contain values of the treatment variable lower that the lower limit of the support of the treatment variable in the status quo allocation. The number of individuals out of support is however relatively small. For example, from allocation 22 (38) onwards, less than 10% (2%) of observations are out of support. These findings have to be taken into account when interpreting reallocation gains.

Inference draws upon the bootstrap (resampling) procedure described in Section 1.5.1. I repeat the estimation for each bootstrap sample. I start by determining the set of 56 allocations for each bootstrap sample according to the reallocation mechanism described in this Section. Then, I implement the plug-in procedure described above in order to compute a new reallocation gain (loss) estimate for each allocation and bootstrap sample separately.

Finding the optimal allocation in terms of average outcomes or in terms of the gender gap in average outcome would also be desirable in this context. Yet, this task is computationally intense, due to the different levels of student quality, and due to the relatively large size of groups as well as the relatively large number of groups. Section 1.C outlines the integer programming problem that has to be solved in order to find the optimal allocation, for each cohort separately. Outlining this problem helps to understand the complexity of the optimization. This paper therefore concentrates on peer and reallocation effects, leaving the derivation of the optimal allocation aside.

1.6 Results

1.6.1 Peer effects

Table 1.2 presents the average partial effects of peer quality on the probability of passing the first year in the first attempt, for the full sample as well as separately for males and females. The table reports average marginal effects across the whole support of peer quality. Additionally, the tables present treatment effects for different sub-samples with respect to an individuals' position in the distribution of peer quality: First, for individuals in the bottom (top) half of the distribution of peer quality, denoted by "Below median" ("Above median"); second, for individuals in each of the four quartiles of the distribution of peer quality.[31] Stratification of the sample in this way allows for detecting non-linearities and threshold effects: For example, peer effects might fade out once individuals reach a certain threshold of predicted peer quality. The table furthermore presents p-values of the randomization inference. For corresponding confidence intervals based on bootstrap inference, I refer to Tables 1.A.6, 1.A.7, and 1.A.8. Significance levels are robust across both types of inference methods.

The average partial effect for the full sample is never significant at any conventional significance level. The picture changes when looking at different quartiles of the treatment distribution: Positive and significant effects arise both for individuals below the median and for individuals below the bottom quartile of the treatment distribution (these groups obviously overlap). The effect is also economically significant: For example, the average treatment effect without trimming and for individuals below the median amounts to 0.17. An increase in predicted peer quality by one standard deviation (0.13) increases the probability of passing by on average 2.2 percentage points, or by approximately 10% of a standard deviation of the outcome.

These results are mainly driven by male students, as separating the results by gender reveals. While male students in the 2 bottom quartiles of the treatment distribution benefit from increases in peer quality, females in the 2 bottom quartiles remain largely unaffected. In contrast, females in the 2 upper quartiles significantly suffer from increases in predicted peer quality, while males in the two upper quartiles experience no significant effect.

The average structural functions and their derivatives visualize a similar pattern (see Figure 1.3). Again, the figures display results for all students as well as for males and females separately. All figures suggest decreasing returns to peer quality, i.e. positive effects across lower quantiles, and no or even negative effects across higher

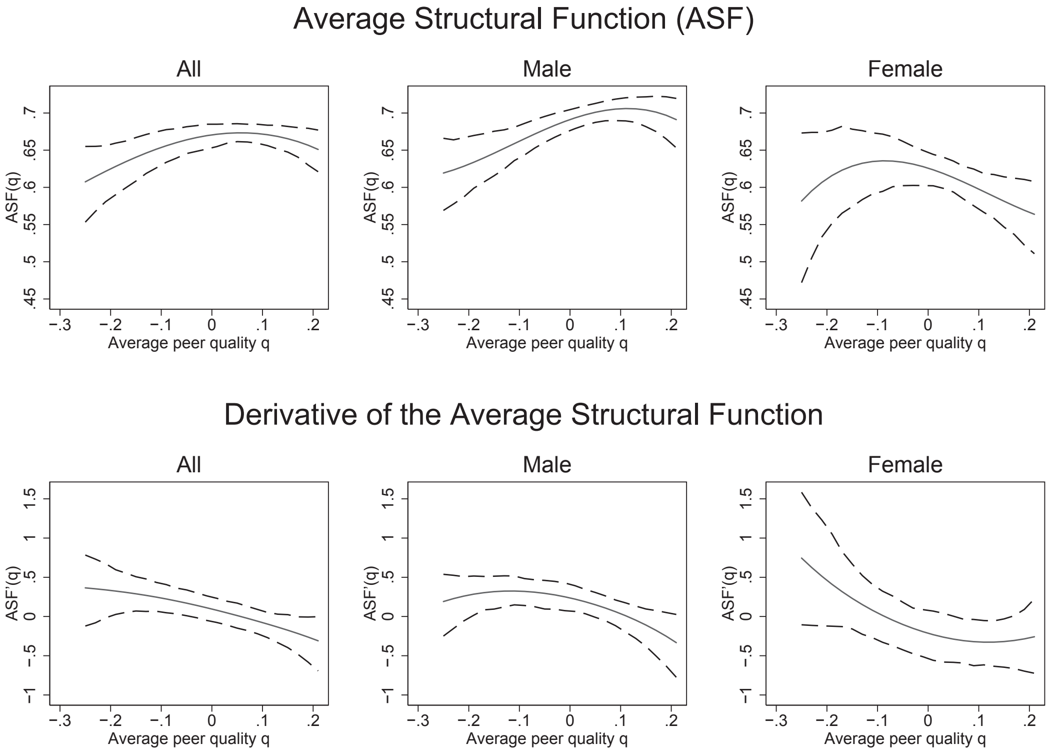
[31] For detailed information on the distribution of own as well as peer quality, see Figure 1.A.2.

Table 1.2: Average Partial Effects (Randomization inference)

All students (n = 5,024)			
Sample	Avg. Partial Effect	Mean	p-value
All	0.01	0.000	0.840
Below median	0.17**	-0.004	0.009
Above median	-0.15	0.003	0.129
Quartile 1	0.25**	0.001	0.008
Quartile 2	0.09	-0.008	0.221
Quartile 3	-0.05	-0.004	0.538
Quartile 4	-0.26*	0.010	0.073
Male students (n = 3,443			
Sample	Avg. Partial Effect	Mean	p-value
All	0.10	-0.010	0.126
Below median	0.25**	-0.008	0.009
Above median	-0.07	-0.012	0.594
Quartile 1	0.27**	-0.007	0.028
Quartile 2	0.24**	-0.010	0.009
Quartile 3	0.09	-0.010	0.226
Quartile 4	-0.24	-0.014	0.193
Female students (n = 1,581)			
Sample	Avg. Partial Effect	Mean	p-value
All	-0.17*	0.019	0.071
Below median	-0.02	0.008	0.868
Above median	-0.30*	0.028	0.051
Quartile 1	0.21	0.021	0.290
Quartile 2	-0.23	-0.004	0.203
Quartile 3	-0.33**	0.007	0.024
Quartile 4	-0.28	0.048	0.304

Average partial effects for the whole samle as well as separated by gender. Samples are specified according to the position of the individual in the distribution of peer quality. Below median: below the median of the peer quality distribution. Above median: Above the median of the peer quality distribution. Quartile 1 (2, 3, 4): In the first (second, third, forth) quartile of the peer quality distribution. The table presents p-values from randomization inference. Column “Estimate” refers to the estimate derived for the status quo allocation. Column “Mean” refers to the mean of the randomization distribution. Inference is based on 250 randomized assignments. ** p<0.05, * p<0.1.

Figure 1.3: Average Structural Function of peer quality



Average structural functions (upper panels) and their derivatives (lower panels). The functions and their derivatives are computed using a grid for peer quality q: For each value of q in the support of peer quality, I compute the probability of passing for each individual and average over all individuals. The model used for computation of the average outcomes control for year dummies and allow for gender heterogeneity. Confidence intervals are based on 250 bootstrap replications. The dashed lines indicate 90% confidence intervals.

quantiles. As the sample consists of fewer females than males, confidence intervals are substantially smaller for males, compared to females.

1.6.2 Reallocation effects

Tables 1.A.10, 1.A.11, and 1.A.12 present the results on reallocation gains, again for all students as well as for males and females separately. The tables show only effects from allocation 22 onwards, as the first 21 rely on predictions out of support of peer quality for more than 10% of the sample (see Section 1.5.2). For the full sample, the analysis predicts weakly significant reallocation losses, compared to the status quo (Table 1.A.10). Losses can be substantial: When switching from the status quo to allocation 22, for example, which is a highly segregated allocation, the average probability of passing declines by only 3.4 percentage points. Female students drive this result: Allocation 22 corresponds to an average reallocation loss for females by 7.4 percentage points, while male students remain largely unaffected. Overall, compared to the set of allocations under study, the status quo allocation performs almost as good as the best of the proposed reallocations.

Figure 1.A.3 illustrates these results further by plotting average outcomes and reallocation gains against the number of the allocation. The figure also displays 90% confidence intervals. The figure illustrates that female students suffer more on average under strongly segregated allocations than males. Moreover, the figure shows that estimates of reallocation effects become increasingly imprecise when moving away from a highly integrated allocation. This is due to the high sampling variability in the fraction of extreme groups (i.e. groups with very high or very low peer quality) across the different bootstrap replications.

Segregation also aggravates the gender gap in the average outcome (see Table 1.A.13). Moving from the status quo allocation to allocation 31, for example, which is more segregated, increases the gender gap in outcomes by 4.5 percentage points on average. Again, comparing the set of reallocations with the status quo, none of the reallocations provides an improvement over the status quo (see also Figure 1.A.4).

To sum up, the magnitude of reallocation effects can be substantial and economically important. Mixing students to achieve an integrated allocation seems a good idea, given the set of allocations under study. This result might be policy relevant if we consider what would have happened if groups were not composed by a randomization device. For example, what would have happened if students were to choose their groups by themselves instead? Answering this question can be an avenue for further research. In particular, studying whether students tend toward more segregated allocations when they are free too choose their groups could speak in favor or against interventions at the onset of university education.

1.6.3 Robustness with respect to trimming

I check for robustness of the results on peer effects with respect to trimming, i.e. I compute the average partial effects under two different trimming rules. According to the first (second) trimming rule, I trim 2.5% (5%) of observations on either side of the distribution of peer quality and compute average partial effects for the trimmed sample. Trimming occurs after the coefficients from Model 1.3 have been derived, and ensures that outliers in terms of peer quality do not drive the average partial effects (or the local average response, respectively). Trimming can induce an increase in precision, but might as well lead to biases. Therefore, comparing estimates under different trimming rules delivers information on the robustness of the results.

Tables 1.A.6, 1.A.7, and 1.A.8 shows average marginal effects and local average responses under no trimming as well as for the two different trimming rules, for the whole sample (Table 1.A.6) as well as for males and females separately (Tables 1.A.7 and 1.A.8, respectively). Moreover, the tables display confidence intervals based on bootstrap inference as well as p-values based on randomization inference. Overall, a robust picture arises from a comparison across trimming rules and inference procedures.

1.7 Discussion

This study investigates peer effects both in a novel setting, i.e. an introductory week and therefore a relatively short intervention, compared to other settings exploited in the literature, and with a novel methodology, i.e. combining several characteristics into a single score. This section compares the findings to findings from the existing literature. Despite differences in setting and methodology, the results seem broadly consistent with existing studies. Signs and magnitudes of the effects appear reasonable. Most of these studies I discuss below investigate ability peer effects, using either high school grade point averages (HSGPA) or SAT scores as ability measures.

For the full sample, as well as for the samples of males and females separately, this paper finds no significant peer effects without further splitting these samples into sub-samples. In this respect, the findings support the studies by Lyle (2007) and Foster (2006). Both authors do not find robust ability peer effects. The settings studied by Lyle (2007) and Foster (2006) compare to the setting presented in this paper, as peer groups are relatively large with 35 and 30 students, respectively. Lyle (2007) investigates military companies with 35 freshmen ("plebes"). Even though interactions between plebes in the same company are intense, especially during the first 6 week at the US Military Academy, no ability spillovers on students' academic performance exist. Instead, Lyle (2007) finds an impact of peers' inclination to stay in the army before entering and students' actual decision to remain in the army for more than 6 years. This finding suggest the relative importance of peer effects on choice outcomes as opposed to ability spillovers. Moreover, Lyle (2007) does not study nonlinear peer effects, which might be an additional reason for his lack of evidence on ability peer effects. Foster (2006) studies peer groups of 30 students that share the same wing and floor in a dormitory. She does not find robust peer effects, even when splitting the sample according to gender. In her interpretation of the findings, her study casts doubt in general on the existence of peer effects through social tie formation. Another explanation for the absence of ability peer effects might be measurement error in academic ability when measured through HSGPA or SAT scores, as discussed by the author. To sum up, both the studies by Lyle (2007) and Foster (2006) both find no ability peer effects, but also suggest using alternative treatment and outcome measures. Departing from ability peer effects, and using a choice-related outcome, this study builds upon these insights.

When splitting the sample according to quartiles of predicted peer quality, the analysis in this paper reveals substantial non-linearities in peer effects as well as peer effects for the students in the bottom quantiles of peer quality. For example, for students at the bottom two quartiles of peer quality, increasing peer quality by 1 standard deviation corresponds to an increase in own probability of passing by 2.2 percentage points, or 0.1 standard deviations. The size of this effect resembles the effect size found by Carrell et al. (2009) in their study on cohorts of the US Air Force Academy: An increase in freshmen SAT by 1 standard deviation corresponds to an increase in freshmen GPA by 0.08 standard deviations. The result of Carrell et al. (2009) may be less strong as they estimate the result over the whole range of peer GPA. Altough their setting is similar to Lyle (2007), one difference seems important. At the US Air Force Academy, freshmen in the same squadron attend the same classes, whereas in the US Military Academy, all plebes in a squadron have the same courses, but do not attend the same classes. If peer effects are determined through study partnerships, this might explain a difference in the results. In the setting studied in this paper, I argue that freshmen peer effects might be mitigated by both friendships and study partnerships that emerge during the first week.

The previously cited studies by Lyle (2007) as well as Carrell et al. (2009) do not investigate effect heterogeneity by gender, supposedly due to the low share of women in the Air Force Academy and the US Military Academy. But some of the roommate studies do, in particular Zimmerman (2003) and Stinebrickner and Stinebrickner (2006). Zimmerman (2003) finds that males' academic outcomes suffer from having a roommate in the lowest 15% of verbal SAT scores. The result does not hold for females, and even reverses when looking at math SAT scores: Females in the bottom 85% of the SAT distribution significantly benefit from having a roommate in the lowest 15% of the SAT math distribution. This result is strikingly in line with the results found in this paper: Peers with low quality, seem harmful to males, but beneficial to females. Existing studies do not explain why these heterogeneities exist, which provides an avenue for further research.[32] Stinebrickner and Stinebrickner (2006) also investigate effect heterogeneity, but their results are different to the results of Zimmerman (2003), and therefore also to the results of this study. While females' academic performance rises in roommates' HSGPA, male students' academic performance remains unaffected. Similar to Stinebrickner and Stinebrickner (2006), Arcidiacono and Nicholson (2005) find positive ability peer effects only for female students. Due to contradictory finding in these different studies, further investigations into gender heterogeneity seems worthwhile.

Reallocation effects from this study are in line with the results by Lyle (2009), who finds positive effects of integration in terms of ability on students' academic performance. In his study, the author uses the same setting as in his previous study on ability peer effects (Lyle, 2007). An increase in the 75–25 differential in math SAT scores corresponds to an increase in average GPA by 0.16 standard deviations. Thus, a reallocation of groups in order to create higher within-group heterogeneity supposedly leads to an increase in average outcomes. Results from the simulation study presented in this paper are consistent with this finding: Allocations with higher within-group variance correspond to higher average outcomes. In contrast to Lyle (2007) however, the model used in this paper does not allow for within-group variance per se to have an effect. Including, for example, within-group variance of peer quality or the 75-25 differential in peer quality as into the model presented here could therefore be a valuable avenue for further research.

Finally, this study emphasizes the persistence of initial contacts at university for subsequent outcomes that occur only one year after the intervention has finished. Only few studies so far investigate whether peer group composition has persistent effects. Sacerdote (2001) finds no persistent effects of roommates' academic ability on students' outcomes after the freshmen year. As discussed above, Lyle (2007) finds an effect of peers' attitude toward the army on own probability of remaining in the army after 6 years, therefore supporting the hypothesis of a long-term impact.

[32] Research in behavioral economics may help explain why these gender specific patterns emerge, see for example Niederle and Vesterlund (2007) on females' reluctance to enter competitive environments.

Similarly, Carrell et al. (2009) find persistence of positive peer effects of peers' verbal SAT on students' GPA during the senior year. The results of Lyle (2007) and Carrell et al. (2009) however do not allow for distinguishing the resons for persistence: Either peer groups are persistent, and therefore peers have a persistent influence on academic outcomes, or initial peer groups generate initial shocks on academic outcomes that remain relevant for a long time. Further exploration of the reasons for persistence seem useful in order to better understand the underlying mechanism. Survey results presented in this study support speak in favor of the persistence of initial peer groups.

1.8 Conclusion

This paper derives peer and reallocation effects in higher education, based on a measure of peer quality. Following Pinto (2011), a student's (own) quality is defined as a weighted average of multiple characteristics (e.g. gender, age). Averaging student quality over a student's peer group (excluding himself) results in a measure of peer quality. Using this measure instead of multiple different peer characteristics simplifies the analysis of both peer and reallocation effects. I use data from 6 cohorts of freshmen at the University of St. Gallen to infer peer and reallocation effects in higher education, specifically studying the probability of passing the first year. Motivated by differences in performance between male and female students – male students perform significantly better on average – I also examine effect heterogeneity with respect to gender as well as effects on the gender gap in educational outcomes.

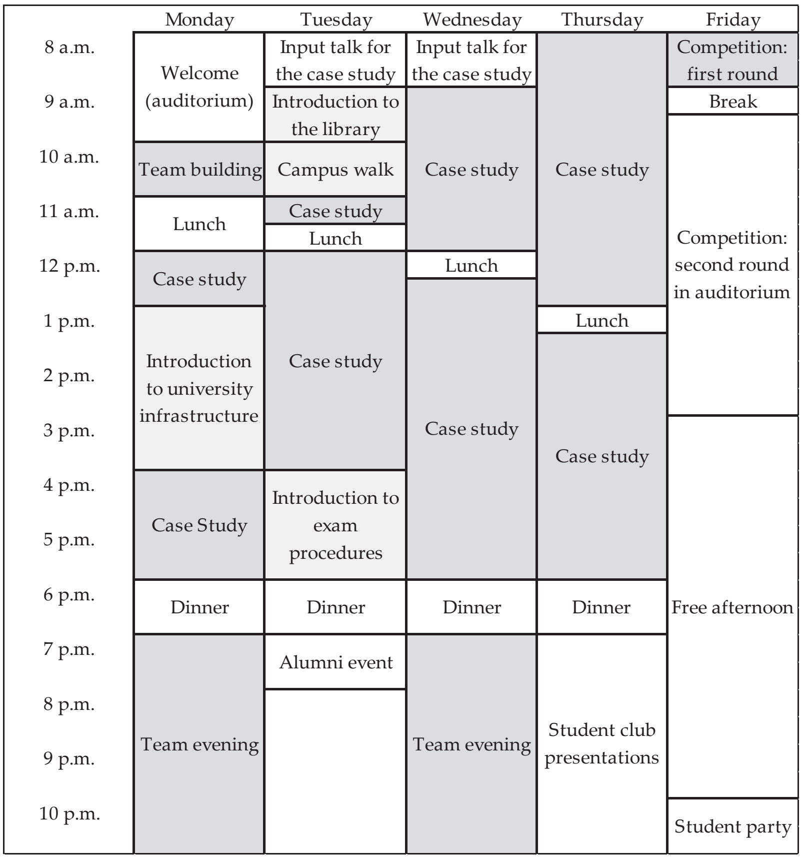
The findings are as follows: First, I find significantly positive effects of an increase in predicted peer quality on academic performance for individuals in the bottom quartile of the distribution of predicted peer quality. This result is largely driven by male students. Second, female students in the upper quartiles of the peer quality distribution react negatively to increases in peer quality. Results 1 and 2 are suggestive of the existence of reallocation effects. Third, analyzing 56 hypothetical reallocations, I derive results on reallocation gains. An increase in segregation seems to induce a decrease in average outcomes, but the effects are only weakly significant.

Forth, segregation aggravates the gender gap in academic outcomes. The results for reallocation gains have to be interpreted with caution: Many of the estimates for highly segregated allocations rely on predictions out of support. Fifth, the status quo allocation, which is close to a fully integrated allocation in terms of within- and between-group variance in the treatment variable, appears close to optimal, both with respect to individual outcomes and with respect to the gender gap.

1.A Figures and tables

Figures

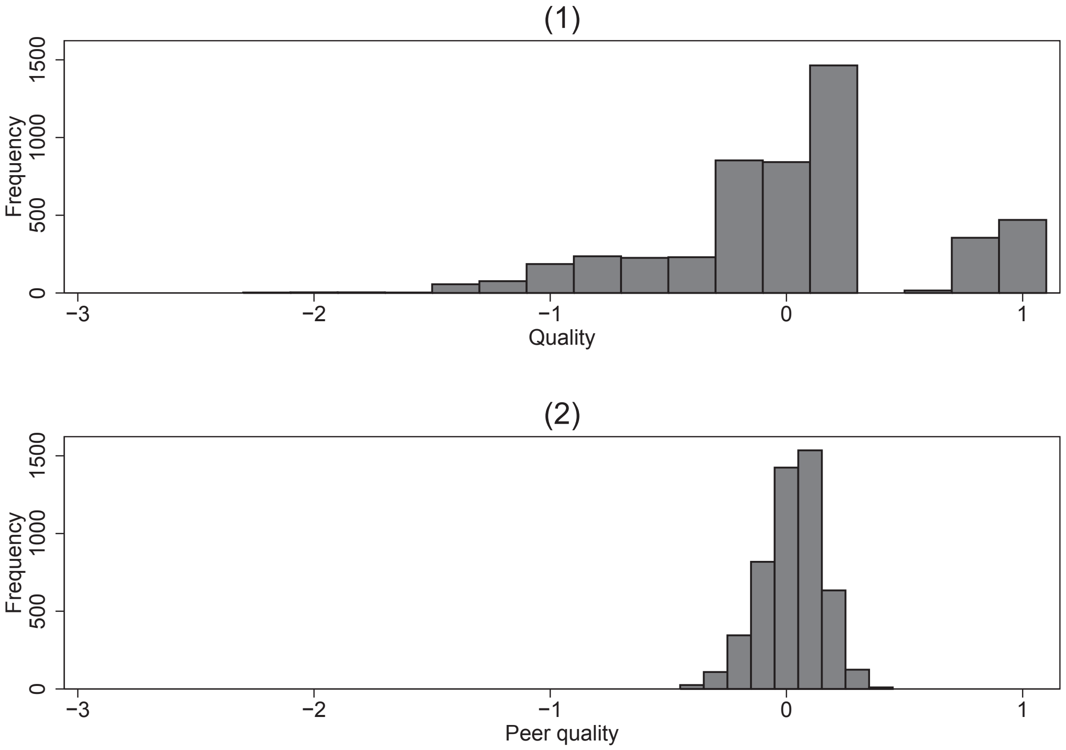
Figure 1.A.1: Schedule of a typical freshmen group



	Monday	Tuesday	Wednesday	Thursday	Friday
8 a.m.	Welcome (auditorium)	Input talk for the case study	Input talk for the case study	Case study	Competition: first round
9 a.m.		Introduction to the library	Case study		Break Competition: second round in auditorium
10 a.m.	Team building	Campus walk			
11 a.m.	Lunch	Case study Lunch			
12 p.m.	Case study	Case study	Lunch Case study		
1 p.m.	Introduction to university infrastructure			Lunch Case study	
2 p.m.					
3 p.m.					Free afternoon
4 p.m.	Case Study	Introduction to exam procedures			
5 p.m.					
6 p.m.	Dinner	Dinner	Dinner	Dinner	
7 p.m.	Team evening	Alumni event	Team evening	Student club presentations	
8 p.m.					
9 p.m.					
10 p.m.					Student party

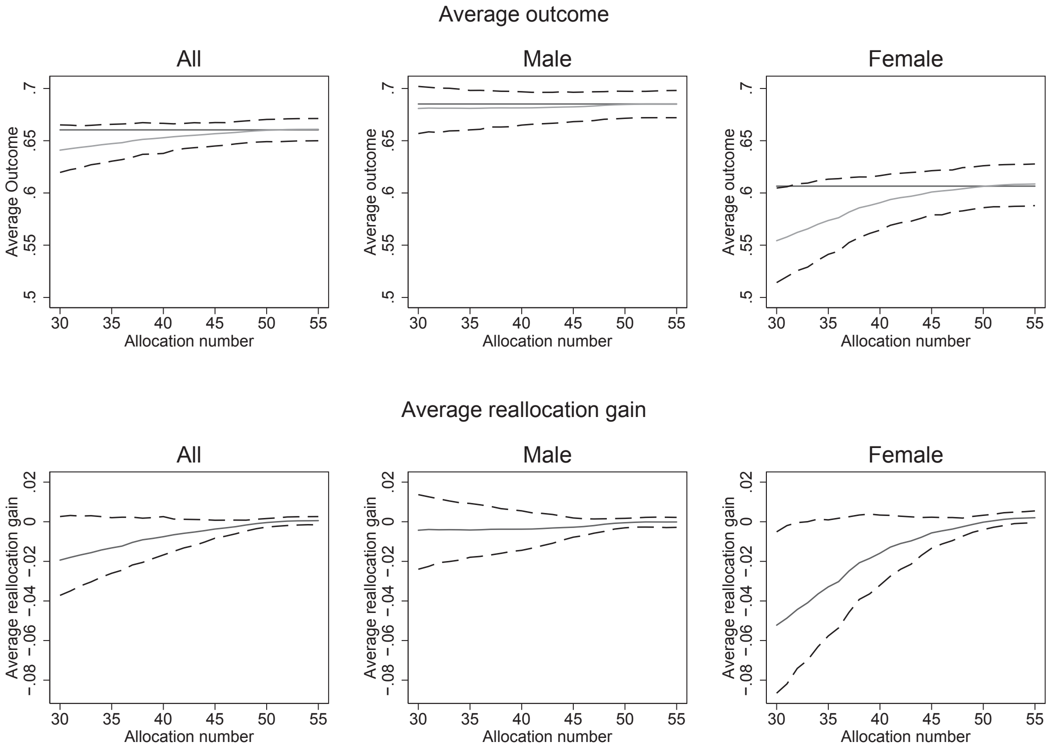
Freshmen week schedule of a typical freshmen group. Dark grey areas indicate time slots spent only in the assigned group, light grey areas indicate time slots spent in assigned groups, but possibly together with other groups, white areas indicate time slots spent not necessarily in assigned groups.

Figure 1.A.2: Distribution of own quality and peer quality



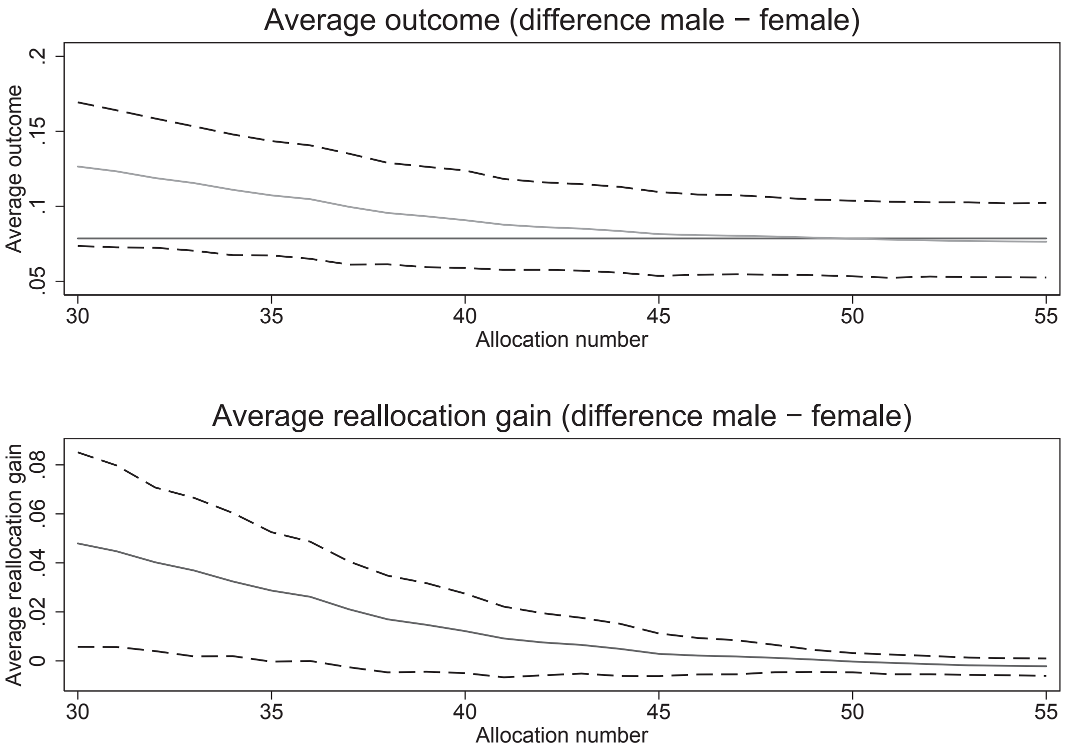
Distribution of own quality (panel 1) as well as peer quality (panel (2)). Predictions for both variables are derived from the logit model presented in section 1.5.1. Peer quality has a mean of 0.04 and a standard deviation of 0.13. 25th percentile (median, 75th percentile): Value of peer quality amounts to -0.06 (0.03, 0.11). Based on 5,024 observations.

Figure 1.A.3: Reallocation effects



Average outcomes and effects of reallocations. Upper panels show average predicted outcomes for hypothetical reallocations first as well as the predicted outcomes for the status quo allocation (horizontal lines). Lower panels show average reallocation gains, compared to the status quo allocation. The x-axis shows the number of the reallocation, according to the assignment mechanism described in Sections 1.5.2 and 1.C. Prediction is based on the full sample (5,024 observations). The models control for year dummies. Confidence intervals based on 250 bootstrap replications. The dashed lines indicate 90% confidence intervals.

Figure 1.A.4: Reallocation effects: Gender gap in outcomes



Average outcomes and effects of reallocations. The upper panel shows average gender gap in predicted outcomes for hypothetical reallocations as well as the predicted outcomes for the status quo allocation (horizontal line). The lower panel shows average reallocation "gains" in the gender difference, compared to the status quo allocation. The x-axis shows the number of the reallocation, according to the assignment mechanism described in Sections 1.5.2 and 1.C. Prediction is based on the full sample (5,024 observations). The models control for year dummies. Confidence intervals based on 250 bootstrap replications. The dashed lines indicate 90% confidence intervals.

Tables

Table 1.A.1: Sample selection: % of freshmen included into the estimation sample

Cohorts	Freshmen	Sample	% of freshmen in sample
All cohorts	5,204	5,024	97%
2003	699	596	85%
2004	636	631	99%
2006	812	806	99%
2007	905	877	97%
2008	1,102	1,082	98%
2009	1,050	1,032	98%

The table illustrates the difference in sample size between the population of entering freshmen and the estimation sample. Sample selection criteria are outlined in Section 1.4.

Table 1.A.2: Descriptive statistics: Number of groups per cohort and group size

Cohort	Obs.	No. of groups	Group size			
			Mean	Median	Min.	Max.
All cohorts	5,024	346	15	15	7	22
2003	596	56	11	11	8	12
2004	631	56	11	12	7	13
2006	806	57	14	14	10	16
2007	877	57	15	15	12	18
2008	1,082	60	18	18	15	22
2009	1,032	60	17	17	11	21

Table 1.A.3: Descriptive statistics by gender

	(1)	(2)	(3)	(4)
	Mean		Difference	t-stat.
	Female	Male		
Pre-treatment characteristics				
Entrance exam	0.12	0.21	-0.09	-7.88
Age (years)	20.01	20.30	-0.29	-4.93
Non-German mother tongue	0.12	0.10	0.02	2.17
Legal studies track	0.12	0.05	0.07	9.26
Foreign nationality	0.19	0.27	-0.08	-6.15
Foreign high school degree	0.17	0.25	-0.09	-6.71
Extended track	0.06	0.05	0.01	1.53
Outcome				
Voluntary dropout during 1st semester	0.08	0.07	0.01	1.85
Failed during 1st semester	0.13	0.10	0.03	3.33
Voluntary dropout during 2nd semester	0.08	0.06	0.02	3.19
Failed during 2nd semester	0.10	0.09	0.01	0.90
First year passed successfully	0.61	0.68	-0.08	-5.46

Descriptive statistics by gender (1,581 female and 3,443 male students) based on administrative records. Column 4 shows the value t-statistics for a mean comparison between males and females (two-sample t-test).

Table 1.A.4: Correlogram: Statistical dependencies between student background variables

	Male	Entrance exam	Age (years)	Foreign mother tongue	Foreign nationality	Foreign high school degree	Legal studies track
Entrance exam	0.111 (0.000)						
Age (years)	0.070 (0.000)	-0.066 (0.000)					
Foreign mother tongue	-0.031 (0.030)	-0.047 (0.001)	-0.089 (0.000)				
Foreign nationality	0.087 (0.000)	0.836 (0.000)	-0.072 (0.000)	-0.003 (0.826)			
Foreign high school degree	0.094 (0.000)	0.872 (0.000)	-0.068 (0.000)	-0.012 (0.401)	0.711 (0.000)		
Legal studies track	-0.130 (0.000)	-0.102 (0.000)	0.111 (0.000)	-0.049 (0.001)	-0.081 (0.000)	-0.103 (0.000)	
Extended track	-0.022 (0.126)	-0.076 (0.000)	-0.074 (0.000)	0.695 (0.000)	-0.064 (0.000)	-0.051 (0.000)	-0.046 (0.001)

The table shows correlation coefficients (pairwise correlations) between student background characteristics. P-values in parentheses.

Table 1.A.5: Test for random assignment within strata

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable: Age				Dependent variable: Non-German mother tongue			
Strata	S1	S2	S3	S4	S1	S2	S3	S4
Share male	-0.106	-0.550	0.408	3.693	-0.106	0.551	-1.144	-0.545
	(0.577)	(1.031)	(1.279)	(3.337)	(0.934)	(1.330)	(2.046)	(5.072)
Share with entrance test	-0.356	-0.278	-0.880	5.777	0.435	0.389	1.969	10.368***
	(0.542)	(0.986)	(1.284)	(3.679)	(0.936)	(1.188)	(2.408)	(3.772)
Share age 20+	-0.032	0.663	-0.119	-1.774	-0.272	0.014	-0.814	0.049
	(0.378)	(0.601)	(0.705)	(1.487)	(0.489)	(0.676)	(1.273)	(1.655)
Share non-german mother tongue	0.294	-0.674	-1.615	-0.767	0.062	-0.510	0.876	2.289
	(0.479)	(0.795)	(1.049)	(2.839)	(0.862)	(1.223)	(1.568)	(2.720)
Share legal studies track	0.587	-0.583	-1.335	-1.822	-1.744*	1.196	-2.556	1.823
	(0.556)	(0.934)	(0.932)	(3.239)	(1.002)	(1.100)	(2.334)	(2.515)
Constant	-0.369	-0.894	-0.722	-5.735**	-1.864***	-2.402**	-1.792	-3.704
	(0.420)	(0.828)	(0.944)	(2.565)	(0.698)	(1.025)	(1.579)	(3.169)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R2	0.002	0.016	0.022	0.085	0.009	0.007	0.015	0.085
Number of observations	2,712	1,391	731	190	2,712	1,391	731	190
χ^2-test: Joint significance of peer variables (p-value)	0.864	0.673	0.367	0.303	0.548	0.875	0.593	0.064

*** p<0.01, ** p<0.05, * p<0.1

Logistic regression of individual characteristics (age, non-german mother tongue) on peer characteristics within each of the 4 strata (S1: Male, no entrance test, S2: Female, no entrance test, S3: Male, entrance test, S4: Male, no entrance test). Based on 5,024 observations.

Table 1.A.6: Average Partial Effects of peer quality on the probability of passing – all students

Trimming	Treatment group		Bootstrap Inference						Randomization Inference		
			BS mean	95% CI		90% CI			Estimate	Mean	p-value
No trimming	All		0.01	-0.08	0.11	-0.06	0.10		0.01	0.000	0.840
	Below median	**	0.17	0.03	0.31	0.05	0.29	**	0.17	-0.004	0.009
	Above median		-0.14	-0.32	0.01	-0.30	0.00		-0.15	0.003	0.129
	Quartile 1	**	0.24	0.05	0.41	0.08	0.40	**	0.25	0.001	0.008
	Quartile 2		0.10	-0.07	0.25	-0.05	0.23		0.09	-0.008	0.221
	Quartile 3		-0.04	-0.19	0.10	-0.17	0.08		-0.05	-0.004	0.538
	Quartile 4		-0.25	-0.50	0.03	-0.47	0.00	*	-0.26	0.010	0.073
5%	All		0.02	-0.09	0.14	-0.08	0.12		0.02	-0.003	0.734
	Below median	**	0.16	0.00	0.32	0.03	0.28	*	0.16	-0.007	0.054
	Above median		-0.11	-0.27	0.03	-0.24	0.01		-0.12	0.002	0.225
	Quartile 1	**	0.24	0.06	0.42	0.09	0.37	**	0.25	-0.006	0.000
	Quartile 2		0.10	-0.07	0.25	-0.05	0.23		0.09	-0.008	0.221
	Quartile 3		-0.04	-0.19	0.10	-0.17	0.08		-0.05	-0.004	0.538
	Quartile 4	*	-0.20	-0.41	0.01	-0.39	-0.01	*	-0.21	0.009	0.074

Continued on the next page.

Continued from the previous page: Average Partial Effects of peer quality on the probability of passing – all students

Trimming	Treatment group		Bootstrap Inference						Randomization Inference		
			BS mean	95% CI		90% CI			Estimate	Mean	p-value
10%	All		0.02	-0.09	0.12	-0.07	0.11		0.02	-0.002	0.746
	Below median	**	0.17	0.02	0.32	0.04	0.29	**	0.17	-0.006	0.027
	Above median		-0.13	-0.30	0.02	-0.26	0.01		-0.13	0.002	0.156
	Quartile 1	**	0.25	0.07	0.43	0.09	0.39	**	0.26	-0.005	0.000
	Quartile 2		0.10	-0.07	0.25	-0.05	0.23		0.09	-0.008	0.221
	Quartile 3		-0.04	-0.19	0.10	-0.17	0.08		-0.05	-0.004	0.538
	Quartile 4		-0.22	-0.44	0.01	-0.41	-0.02	*	-0.23	0.010	0.075

*** p<0.01, ** p<0.05, * p<0.1. Average partial effects, male students (n = 5,024). Samples are specified according to the position of the individual in the distribution of peer quality. Below median: below the median of the peer quality distribution. Above median: Above the median of the peer quality distribution. Quartile 1 (2, 3, 4): In the first (second, third, forth) quartile of the peer quality distribution. Trimming refers to trimming an equal fraction of observations according to the distribution of predicted peer quality. I.e. "Trimming 5%" means that individuals within the lowest or highest 2.5% of the distribution of predicted peer quality, respectively, are excluded. The table presents inference results based on both bootstrap and randomization inference. Column “Estimate” refers to the estimate derived for the status quo allocation. Column “Mean” refers to the mean of the randomization distribution. Inference is based on 250 bootstrap replications and 250 randomized assignment, respectively.

Table 1.A.7: Average Partial Effects of peer quality on the probability of passing – male students

Trimming	Treatment group		Bootstrap Inference						Randomization Inference		
			BS mean	95% CI		90% CI			Estimate	Mean	p-value
No trimming	All		0.10	-0.04	0.22	0.00	0.20		0.10	-0.010	0.840
	Below median	**	0.25	0.07	0.41	0.11	0.38	**	0.25	-0.008	0.009
	Above median		-0.07	-0.28	0.14	-0.24	0.10		-0.07	-0.012	0.129
	Quartile 1	**	0.26	0.06	0.46	0.09	0.44	**	0.27	-0.007	0.008
	Quartile 2	**	0.23	0.01	0.40	0.09	0.39	**	0.24	-0.010	0.221
	Quartile 3		0.09	-0.08	0.25	-0.04	0.23		0.09	-0.010	0.538
	Quartile 4										0.073
			-0.23	-0.55	0.12	-0.47	0.07		-0.24	-0.014	
5%	All		0.13	-0.03	0.27	0.02	0.25	**	0.13	-0.010	0.734
	Below median	**	0.26	0.04	0.43	0.11	0.40	**	0.27	-0.009	0.054
	Above median		-0.01	-0.20	0.17	-0.17	0.14		-0.01	-0.010	0.225
	Quartile 1	**	0.30	0.09	0.49	0.13	0.46	**	0.31	-0.010	0.000
	Quartile 2	**	0.23	0.01	0.40	0.09	0.39	**	0.24	-0.010	0.221
	Quartile 3		0.09	-0.08	0.25	-0.04	0.23		0.09	-0.010	0.538
	Quartile 4										0.074
			-0.14	-0.38	0.11	-0.34	0.08		-0.15	-0.010	

Continued on the next page.

Continued from the previous page: Average Partial Effects of peer quality on the probability of passing – male students

Trimming	Treatment group		Bootstrap Inference						Randomization Inference		
			BS mean	95% CI		90% CI			Estimate	Mean	p-value
10%	All	*	0.12	-0.03	0.25	0.00	0.23	**	0.12	-0.010	0.746
	Below median	**	0.26	0.06	0.42	0.11	0.40	**	0.27	-0.009	0.027
	Above median		-0.03	-0.24	0.16	-0.20	0.12		-0.04	-0.011	0.156
	Quartile 1	**	0.29	0.08	0.49	0.12	0.46	**	0.30	-0.010	0.000
	Quartile 2	**	0.23	0.01	0.40	0.09	0.39	**	0.24	-0.010	0.221
	Quartile 3		0.09	-0.08	0.25	-0.04	0.23		0.09	-0.010	0.538
	Quartile 4		-0.18	-0.45	0.11	-0.40	0.08		-0.19	-0.011	0.075

*** p<0.01, ** p<0.05, * p<0.1. Average partial effects, male students (n = 3,443). Samples are specified according to the position of the individual in the distribution of peer quality. Below median: below the median of the peer quality distribution. Above median: Above the median of the peer quality distribution. Quartile 1 (2, 3, 4): In the first (second, third, forth) quartile of the peer quality distribution. Trimming refers to trimming an equal fraction of observations according to the distribution of predicted peer quality. I.e. "Trimming 5%" means that individuals within the lowest or highest 2.5% of the distribution of predicted peer quality, respectively, are excluded. The table presents inference results based on both bootstrap and randomization inference. Column “Estimate” refers to the estimate derived for the status quo allocation. Column “Mean” refers to the mean of the randomization distribution. Inference is based on 250 bootstrap replications and 250 randomized assignment, respectively.

Table 1.A.8: Average Partial Effects of peer quality on the probability of passing – female students

Trimming	Sample		Bootstrap Inference						Randomization Inference		
			BS mean	95% CI		90% CI			Estimate	Mean	p-value
No trimming	All		-0.17	-0.39	0.03	-0.35	0.01	*	-0.17	0.019	0.840
	Below median		-0.02	-0.30	0.27	-0.26	0.22		-0.02	0.008	0.009
	Above median	**	-0.29	-0.62	-0.02	-0.56	-0.05	*	-0.30	0.028	0.129
	Quartile 1		0.20	-0.24	0.54	-0.15	0.50		0.21	0.021	0.008
	Quartile 2		-0.20	-0.57	0.13	-0.52	0.07		-0.23	-0.004	0.221
	Quartile 3	*	-0.31	-0.64	0.00	-0.56	-0.07	**	-0.33	0.007	0.538
	Quartile 4		-0.28	-0.77	0.15	-0.68	0.07		-0.28	0.048	0.073
5%	All		-0.20	-0.46	0.03	-0.41	0.00	*	-0.21	0.012	0.734
	Below median		-0.08	-0.40	0.21	-0.35	0.15		-0.09	0.000	0.054
	Above median	**	-0.31	-0.62	-0.05	-0.58	-0.08	**	-0.33	0.022	0.225
	Quartile 1		0.09	-0.24	0.42	-0.19	0.36		0.10	0.005	0.000
	Quartile 2		-0.20	-0.57	0.13	-0.52	0.07		-0.23	-0.004	0.221
	Quartile 3	*	-0.31	-0.64	0.00	-0.56	-0.07	**	-0.33	0.007	0.538
	Quartile 4	**	-0.32	-0.70	0.01	-0.64	-0.03	*	-0.33	0.040	0.074

Continued on the next page.

Continued from the previous page: Average Partial Effects of peer quality on the probability of passing – female students

Trimming	Sample		Bootstrap Inference						Randomization Inference		
			BS mean	95% CI		90% CI			Estimate	Mean	p-value
10%	All		-0.19	-0.44	0.03	-0.38	0.00	*	-0.20	0.015	0.746
	Below median		-0.05	-0.36	0.23	-0.31	0.18		-0.06	0.002	0.027
	Above median	**	-0.31	-0.60	-0.04	-0.58	-0.07	**	-0.32	0.025	0.156
	Quartile 1		0.14	-0.24	0.47	-0.18	0.44		0.15	0.009	0.000
	Quartile 2		-0.20	-0.57	0.13	-0.52	0.07		-0.23	-0.004	0.221
	Quartile 3	*	-0.31	-0.64	0.00	-0.56	-0.07	*	-0.33	0.007	0.538
	Quartile 4		-0.31	-0.71	0.03	-0.64	0.00		-0.32	0.043	0.075

*** p<0.01, ** p<0.05, * p<0.1. Average partial effects, female students (n = 1,581). Samples are specified according to the position of the individual in the distribution of peer quality. Below median: below the median of the peer quality distribution. Above median: Above the median of the peer quality distribution. Quartile 1 (2, 3, 4): In the first (second, third, forth) quartile of the peer quality distribution. Trimming refers to trimming an equal fraction of observations according to the distribution of predicted peer quality. I.e. "Trimming 5%" means that individuals within the lowest or highest 2.5% of the distribution of predicted peer quality, respectively, are excluded. The table presents inference results based on both bootstrap and randomization inference. Column “Estimate” refers to the estimate derived for the status quo allocation. Column “Mean” refers to the mean of the randomization distribution. Inference is based on 250 bootstrap replications and 250 randomized assignment, respectively.

Table 1.A.9: Hypothetical allocations: Values of peer quality outside the support of peer quality in the status quo allocation

Allo- cation number	Min. of peer quality (pq)	% (#) individuals w/ out-of-sample prediction (pq $<$ 0.5394)		Max. of peer quality (pq)	% (#) individuals w/ out-of-sample prediction (pq $>$ 0.773)	
1	0.281	20%	(982)	0.895	18%	(881)
2	0.359	18%	(899)	0.895	18%	(881)
3	0.385	16%	(798)	0.895	18%	(881)
4	0.398	14%	(708)	0.895	16%	(828)
5	0.408	13%	(675)	0.895	17%	(833)
6	0.415	12%	(620)	0.895	17%	(833)
7	0.429	10%	(507)	0.895	14%	(685)
8	0.433	9%	(476)	0.881	13%	(658)
9	0.447	8%	(394)	0.881	13%	(649)
10	0.452	7%	(329)	0.881	12%	(615)
11	0.466	5%	(269)	0.874	10%	(486)
12	0.470	4%	(206)	0.855	10%	(492)
13	0.485	3%	(140)	0.855	10%	(492)
14	0.501	2%	(95)	0.855	7%	(368)
15	0.513	1%	(42)	0.855	7%	(337)
16	0.523	1%	(35)	0.847	6%	(309)
17	0.524	0%	(14)	0.841	6%	(310)
18	0.536	0%	(10)	0.841	6%	(283)
19	0.536	0%	(10)	0.841	5%	(254)
20	0.536	0%	(10)	0.829	5%	(231)
21	0.536	0%	(10)	0.829	4%	(201)
22	0.536	0%	(10)	0.829	4%	(180)
23	0.539	0%	(9)	0.829	4%	(180)
24	0.539	0%	(9)	0.817	3%	(151)
25	0.546	0%	(0)	0.817	3%	(136)
26	0.546	0%	(0)	0.817	2%	(125)
27	0.546	0%	(0)	0.817	2%	(116)

Continued on the next page.

Continued from the previous page: Hypothetical allocations: Values of peer quality outside the support of peer quality in the status quo allocation

Allocation number	Min. of peer quality (pq)	% (#) individuals w/ out-of-sample prediction (pq $<$ 0.5394)		Max. of peer quality (pq)	% (#) individuals w/ out-of-sample prediction (pq $>$ 0.773)	
28	0.546	0%	(0)	0.817	2%	(92)
29	0.548	0%	(0)	0.806	2%	(88)
30	0.548	0%	(0)	0.806	2%	(82)
31	0.548	0%	(0)	0.806	2%	(80)
32	0.548	0%	(0)	0.806	1%	(46)
33	0.548	0%	(0)	0.795	1%	(35)
34	0.548	0%	(0)	0.795	1%	(35)
35	0.552	0%	(0)	0.795	1%	(35)
36	0.553	0%	(0)	0.795	1%	(28)
37	0.553	0%	(0)	0.785	0%	(13)
38	0.553	0%	(0)	0.785	0%	(13)
39	0.553	0%	(0)	0.785	0%	(13)
40	0.553	0%	(0)	0.785	0%	(13)
41	0.553	0%	(0)	0.775	0%	(2)
42	0.553	0%	(0)	0.775	0%	(2)
43	0.553	0%	(0)	0.775	0%	(2)
44	0.553	0%	(0)	0.775	0%	(2)
45	0.553	0%	(0)	0.768	0%	(0)
46	0.553	0%	(0)	0.768	0%	(0)
47	0.553	0%	(0)	0.768	0%	(0)
48	0.553	0%	(0)	0.766	0%	(0)
49	0.553	0%	(0)	0.760	0%	(0)
50	0.553	0%	(0)	0.760	0%	(0)
51	0.553	0%	(0)	0.760	0%	(0)
52	0.553	0%	(0)	0.759	0%	(0)
53	0.553	0%	(0)	0.759	0%	(0)
54	0.553	0%	(0)	0.759	0%	(0)

Table 1.A.10: Reallocation gains: All students

Allocation number		Average outcome	Difference to Status quo (SQ = 0.660)	95% CI		90% CI	
22	*	0.626	-0.034	-0.065	0.002	-0.061	-0.003
23	*	0.628	-0.032	-0.061	0.004	-0.057	-0.001
24		0.631	-0.029	-0.056	0.004	-0.051	0.000
25		0.633	-0.027	-0.052	0.007	-0.048	0.001
26		0.635	-0.025	-0.051	0.008	-0.045	0.001
27		0.637	-0.023	-0.047	0.009	-0.044	0.003
28		0.639	-0.021	-0.044	0.009	-0.040	0.003
29		0.640	-0.020	-0.042	0.008	-0.038	0.003
30		0.641	-0.019	-0.040	0.009	-0.037	0.003
31		0.642	-0.018	-0.036	0.009	-0.035	0.003
32		0.644	-0.017	-0.034	0.008	-0.032	0.003
33		0.645	-0.016	-0.033	0.007	-0.030	0.003
34		0.646	-0.014	-0.032	0.007	-0.028	0.003
35		0.647	-0.013	-0.029	0.007	-0.026	0.002
36		0.648	-0.012	-0.027	0.007	-0.024	0.002
37		0.650	-0.010	-0.023	0.006	-0.022	0.002
38		0.651	-0.009	-0.022	0.006	-0.021	0.002
39		0.652	-0.008	-0.021	0.005	-0.019	0.002
40		0.653	-0.008	-0.018	0.005	-0.017	0.003
41		0.654	-0.007	-0.016	0.004	-0.015	0.001
42		0.655	-0.006	-0.014	0.004	-0.013	0.001
43		0.655	-0.005	-0.013	0.003	-0.012	0.001
44		0.656	-0.004	-0.011	0.002	-0.010	0.001
45		0.657	-0.004	-0.009	0.001	-0.008	0.001
46		0.657	-0.003	-0.008	0.001	-0.007	0.001
47		0.658	-0.002	-0.007	0.002	-0.006	0.001
48		0.659	-0.002	-0.005	0.002	-0.005	0.001
49		0.659	-0.001	-0.004	0.002	-0.004	0.001
50		0.660	0.000	-0.003	0.002	-0.003	0.002
51		0.660	0.000	-0.003	0.002	-0.002	0.002
52		0.661	0.000	-0.002	0.003	-0.002	0.002
53		0.661	0.000	-0.002	0.003	-0.002	0.003
54		0.661	0.000	-0.002	0.003	-0.001	0.003
55		0.661	0.001	-0.002	0.003	-0.002	0.003

*** p<0.01, ** p<0.05, * p<0.1. The table presents bootstrap means and confidence intervals for each allocation, starting from allocation 22 (less than 10% of observations out of support). Based on 250 bootstrap replications. See Section 1.5.2.

Table 1.A.11: Reallocation gains: Male students

Allocation	Average	Difference	95% CI		90% CI	
number	outcome	to Status quo				
		(SQ = 0.685)				
22	0.669	-0.016	-0.054	0.018	-0.046	0.013
23	0.672	-0.014	-0.050	0.020	-0.041	0.014
24	0.674	-0.011	-0.045	0.019	-0.037	0.016
25	0.675	-0.010	-0.040	0.018	-0.034	0.015
26	0.677	-0.008	-0.039	0.017	-0.031	0.015
27	0.679	-0.006	-0.036	0.018	-0.029	0.016
28	0.681	-0.005	-0.033	0.017	-0.026	0.016
29	0.681	-0.004	-0.030	0.016	-0.024	0.014
30	0.681	-0.004	-0.028	0.015	-0.024	0.014
31	0.681	-0.004	-0.028	0.014	-0.023	0.013
32	0.681	-0.004	-0.025	0.013	-0.021	0.012
33	0.681	-0.004	-0.024	0.012	-0.020	0.011
34	0.681	-0.004	-0.023	0.012	-0.019	0.010
35	0.681	-0.004	-0.021	0.010	-0.018	0.009
36	0.681	-0.004	-0.021	0.009	-0.017	0.009
37	0.681	-0.004	-0.019	0.009	-0.017	0.008
38	0.681	-0.004	-0.018	0.008	-0.016	0.007
39	0.681	-0.004	-0.018	0.007	-0.015	0.006
40	0.681	-0.004	-0.016	0.007	-0.014	0.006
41	0.681	-0.004	-0.014	0.005	-0.013	0.005
42	0.682	-0.003	-0.013	0.005	-0.012	0.004
43	0.682	-0.003	-0.012	0.005	-0.011	0.004
44	0.682	-0.003	-0.011	0.003	-0.009	0.003
45	0.682	-0.003	-0.009	0.002	-0.008	0.002
46	0.683	-0.002	-0.008	0.002	-0.007	0.002
47	0.683	-0.002	-0.007	0.002	-0.006	0.001
48	0.684	-0.001	-0.005	0.002	-0.005	0.002
49	0.684	-0.001	-0.004	0.002	-0.004	0.002
50	0.685	0.000	-0.004	0.002	-0.003	0.002
51	0.685	0.000	-0.004	0.002	-0.003	0.002
52	0.685	0.000	-0.003	0.003	-0.003	0.002
53	0.685	0.000	-0.004	0.003	-0.003	0.002
54	0.685	0.000	-0.004	0.003	-0.003	0.002
55	0.685	0.000	-0.004	0.003	-0.003	0.002

*** $p<0.01$, ** $p<0.05$, * $p<0.1$. The table presents bootstrap means and confidence intervals for each allocation, starting from allocation 22 (less than 10% of observations out of support). Based on 250 bootstrap replications. See Section 1.5.2.

Table 1.A.12: Reallocation gains: Female students

Allocation number		Average outcome	Difference to Status quo (SQ = 0.607)	95% CI		90% CI	
22	*	0.533	-0.074	-0.120	0.023	-0.116	-0.003
23	*	0.533	-0.073	-0.118	0.021	-0.115	-0.007
24	*	0.538	-0.068	-0.115	0.018	-0.109	-0.008
25	*	0.542	-0.065	-0.108	0.011	-0.104	-0.009
26	*	0.544	-0.063	-0.106	0.010	-0.103	-0.010
27	*	0.545	-0.061	-0.103	0.008	-0.099	-0.012
28	*	0.549	-0.057	-0.099	0.004	-0.094	-0.011
29	*	0.552	-0.055	-0.095	0.003	-0.091	-0.006
30	*	0.554	-0.052	-0.091	0.002	-0.087	-0.005
31	*	0.558	-0.049	-0.084	0.004	-0.082	-0.002
32		0.562	-0.044	-0.078	0.005	-0.074	0.000
33		0.566	-0.041	-0.074	0.008	-0.070	0.000
34		0.570	-0.036	-0.068	0.010	-0.063	0.001
35		0.574	-0.033	-0.062	0.010	-0.058	0.001
36		0.576	-0.030	-0.056	0.009	-0.054	0.002
37		0.582	-0.025	-0.049	0.010	-0.046	0.003
38		0.586	-0.021	-0.043	0.015	-0.039	0.003
39		0.588	-0.019	-0.040	0.013	-0.036	0.004
40		0.591	-0.016	-0.034	0.013	-0.032	0.003
41		0.594	-0.013	-0.029	0.008	-0.028	0.003
42		0.596	-0.011	-0.025	0.007	-0.024	0.003
43		0.597	-0.010	-0.023	0.006	-0.021	0.002
44		0.599	-0.008	-0.020	0.004	-0.017	0.002
45		0.601	-0.006	-0.015	0.004	-0.013	0.002
46		0.602	-0.005	-0.013	0.004	-0.011	0.002
47		0.603	-0.004	-0.011	0.004	-0.010	0.002
48		0.604	-0.002	-0.009	0.003	-0.007	0.002
49		0.605	-0.001	-0.007	0.003	-0.005	0.003
50		0.606	0.000	-0.005	0.004	-0.004	0.003
51		0.607	0.001	-0.004	0.005	-0.003	0.004
52		0.608	0.001	-0.002	0.006	-0.002	0.005
53		0.608	0.002	-0.001	0.006	-0.001	0.005
54		0.608	0.002	-0.001	0.006	-0.001	0.005
55		0.609	0.002	-0.001	0.006	-0.001	0.005

*** p<0.01, ** p<0.05, * p<0.1. The table presents bootstrap means and confidence intervals for each allocation, starting from allocation 22 (less than 10% of observations out of support). Based on 250 bootstrap replications. See Section 1.5.2.

Table 1.A.13: Reallocation gains: Difference (male - female)

Allocation number		Average outcome	Difference to Status quo (SQ = 0.079)	95% CI		90% CI	
22		0.136	0.057	-0.040	0.115	-0.002	0.112
23	*	0.138	0.060	-0.032	0.117	0.002	0.112
24	*	0.135	0.057	-0.025	0.112	0.007	0.107
25	*	0.133	0.055	-0.013	0.109	0.005	0.101
26	*	0.134	0.055	-0.006	0.106	0.007	0.099
27	*	0.134	0.056	-0.003	0.104	0.008	0.097
28	**	0.131	0.053	0.001	0.100	0.009	0.093
29	**	0.129	0.050	0.001	0.095	0.008	0.089
30	**	0.127	0.048	0.000	0.088	0.006	0.085
31	**	0.123	0.045	0.000	0.084	0.006	0.080
32	*	0.119	0.040	-0.001	0.073	0.004	0.071
33	*	0.116	0.037	-0.001	0.070	0.002	0.067
34	*	0.111	0.032	-0.003	0.063	0.002	0.060
35	*	0.107	0.029	-0.007	0.057	0.000	0.053
36	*	0.105	0.026	-0.007	0.052	0.000	0.049
37		0.100	0.021	-0.008	0.046	-0.003	0.041
38		0.096	0.017	-0.010	0.039	-0.005	0.035
39		0.093	0.015	-0.012	0.035	-0.004	0.032
40		0.091	0.012	-0.011	0.031	-0.005	0.027
41		0.088	0.009	-0.009	0.024	-0.007	0.022
42		0.086	0.008	-0.010	0.022	-0.006	0.019
43		0.085	0.007	-0.010	0.019	-0.005	0.018
44		0.084	0.005	-0.008	0.016	-0.006	0.015
45		0.081	0.003	-0.008	0.013	-0.006	0.011
46		0.081	0.002	-0.008	0.011	-0.006	0.009
47		0.080	0.002	-0.007	0.009	-0.005	0.008
48		0.080	0.001	-0.005	0.008	-0.005	0.007
49		0.079	0.001	-0.005	0.006	-0.004	0.004
50		0.078	0.000	-0.006	0.004	-0.005	0.003
51		0.078	-0.001	-0.006	0.003	-0.005	0.003
52		0.077	-0.001	-0.006	0.002	-0.005	0.002
53		0.077	-0.002	-0.007	0.002	-0.006	0.001
54		0.077	-0.002	-0.007	0.002	-0.006	0.001
55		0.076	-0.002	-0.007	0.001	-0.006	0.001

*** p<0.01, ** p<0.05, * p<0.1. The table presents bootstrap means and confidence intervals for each allocation, starting from allocation 22 (less than 10% of observations out of support). Based on 250 bootstrap replications. See Section 1.5.2.

1.B Questionnaire

This section presents translations of the questions on formation of friendship and study partnerships. The original language of the questionnaire is German.[33]

Did you participate in the freshmen week? (Yes, at least for 1 day. - No.)

The questionnaire proceeds to the following question if the student answers Yes.

Please write down for yourself the 5 individuals with whom you so far spent most of free time during the period of your Bachelor degree (including lunch, coffee breaks, student club activities, sports activities, parties, holidays). Please exclude time periods spent on exchange or on internships.

1. How many of these individuals did you get to know at the University of St. Gallen?
2. How many of these individuals are your freshmen team members?

Please write down for yourself the 5 individuals with whom you so far spent most of your studying activities during the period of your Bachelor degree (e.g. group presentations, learning together, written group work, appointments for studying in the library). Please exclude time periods spent on exchange or on internships.

1. How many of these individuals are your freshmen team members?

Please recall your freshmen team.

1. With how many of your team members did you spent your free time during the last 6 months (including lunch, coffee breaks, student club activities, sports activities, parties, holidays)?
2. With how many of your team members did you spent time studying together during the last 6 months (including group presentations, learning together, written group work, appointments for studying in the library).

[33] The full questionnaire as well as the German version are available upon request form the author.

1.C Technical appendix

Maximization problem for the optimal allocation

Example for 1 cohort with 60 groups and 25 different levels of student quality (i.e. 5 binary characteristics).
The following parameters are given, either by the setup, or by previous estimation of the model:

- Assume we have 60 groups ($G = 60$).
- Potentially 25 types, i.e. 25 levels of student quality ($T = 25$).
- N is the overall number of observations.
- $N_1, ..., N_{25}$ is the number of observations for each type.
- Groupsize is s_g.
- Lower bound of average peer quality is q_{lb} and upper bound is q_{ub} (estimated).
- Quality of each individual of type t is q^t (estimated).

The following parameters have to be determined through optimization:

- Average quality of group g is $\overline{q}_g$.
- Average peer quality for each individual of type t in group g (excluding himself) is $\overline{q}_g^t$.
- Number of individuals with type t in group g is $n_{t,g}$.

An allocation is defined as a 25x60 (TxG) assignment matrix A, with

$$A = \begin{pmatrix} n_{1,1} & \dots & n_{1,60} \\ \vdots & n_{t,g} & \vdots \\ n_{25,1} & \dots & n_{25,60} \end{pmatrix}. \qquad (1.13)$$

Objective function

$$\max_{n_{t,g}} \frac{1}{N} \sum_{g=1}^{G} \sum_{t=1}^{T} n_{t,g} F(\overline{q}_g^t, q^t), \tag{1.14}$$

where $F(\overline{q}_g^t, q^t)$ is the outcome for type t when peer quality is $\overline{q}_g^t$. Specifically,

$$F(\overline{q}_g^t, q^t) = \Lambda\left(q^t + \sum_{k=1}^{K} (\gamma_{1k} + \gamma_{2k}\text{male}_t)(\overline{q}_g^t)^k\right), \tag{1.15}$$

so that the objective function is

$$\max_{n_{t,g}} \frac{1}{N} \sum_{g=1}^{G} \sum_{t=1}^{T} n_{t,g} \Lambda\left(q^t + \sum_{k=1}^{K} (\gamma_{1k} + \gamma_{2k}\text{male}_t)(\overline{q}_g^t)^k\right), \tag{1.16}$$

with $k = 3$ and γ_{1k}, γ_{2k} given.

Equality constraints on the number of individuals per type in the cohort (25 restrictions)

$$n_{1,1} + \cdots + n_{1,60} = N_1 \tag{1.17}$$

$$\vdots \tag{1.18}$$

$$n_{25,1} + \cdots + n_{25,60} = N_{25} \tag{1.19}$$

Equality constraints on group size (60 restrictions)

$$n_{1,1} + \cdots + n_{25,1} = s_1 \tag{1.20}$$

$$\vdots \tag{1.21}$$

$$n_{1,60} + \cdots + n_{25,60} = s_{60} \tag{1.22}$$

Computation of average quality for each group (60 equations)

$$\begin{aligned}
\overline{q}_1 &= \frac{1}{s_1}(n_{1,1}q^1 + \cdots + n_{25,1}q^{25}) && (1.23)\\
&\vdots && (1.24)\\
\overline{q}_{60} &= \frac{1}{n_{60}}(n_{1,60}q^1 + \cdots + n_{25,60}q^{25}) && (1.25)
\end{aligned}$$

Computation of average quality for each individual of type t in each group (60x25 equations)

$$\begin{aligned}
\overline{q}_1^t &= \frac{s_1}{s_1 - 1}\overline{q}_1 - q^t && (1.26)\\
&\vdots && (1.27)\\
\overline{q}_{60}^t &= \frac{s_{60}}{s_{60} - 1}\overline{q}_{60} - q^t && (1.28)
\end{aligned}$$

Inequality constraints on the support of peer quality (120 restrictions)

$$\begin{aligned}
q_1 &\geq q_{lb} && (1.29)\\
&\vdots && (1.30)\\
g_{60} &\geq g_{lb}, && (1.31)
\end{aligned}$$

and

$$\begin{aligned}
q_1 &\leq q_{ub} && (1.32)\\
&\vdots && (1.33)\\
g_{60} &\leq g_{ub}. && (1.34)
\end{aligned}$$

Computation of within- and between-variances

(Average) within-group variance is given as

$$\widehat{\mathbb{V}}^w = \frac{1}{G}\sum_{g=1}^{G}\hat{\sigma}_g^2 = \frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{M_g-1}\sum_{i=1}^{M_g}(\widehat{PQ}_{ig} - \overline{PQ}_g)^2\right), \tag{1.35}$$

where $\widehat{PQ}_{ig}$ denotes (predicted) quality of student i in group g, G denotes the number of groups, M_g denotes the size of group g, and $\overline{PQ}_g = \frac{1}{M_g}\sum_{i=1}^{M_g}\widehat{PQ}_{ig}$.

Between-group variance is defined as

$$\widehat{\mathbb{V}}^b = \frac{1}{G-1}\sum_{g=1}^{G}(\overline{PQ}_g - \overline{PQ})^2, \tag{1.36}$$

where $\overline{PQ} = \frac{1}{G}\sum_{g=1}^{G}\overline{PQ}_g$. As group sizes differ, $\overline{PQ}$ can differ from $\frac{1}{n}\sum_{i=1}^{n}\widehat{PQ}_{ig}$.

Reallocation algorithm

Algorithm to create continuum of allocations ("continuous" shift from segregated to integrated allocations)

- Sort individuals according to their own quality (1 = lowest quality, ..., 6 = highest quality).

- For illustration purposes, suppose that the original sample consists of 3 groups with 2 group members each (the algorithm can easily be extended to more and bigger groups).

- Place the 1st and 2nd student in group 1, the 3rd and 4th student in group 2, ..., until all groups are filled (allocation 1 is fully segregated).

- To move on to the next allocation, group 1 is dissolved. The 1st student is assigned to group 1, the 2nd student to group 2. The remaining students are assigned to the remaining slots according to their rank in the distribution of predicted (own) quality.

- Proceed until all groups are dissolved. I.e., with 3 groups, we end up with 3 allocations.

Table 1.C.14: Allocations (Example)

	Segregated		Integrated
Group #	Allocation 1	Allocation 2	Allocation 3
1	1	1	1
1	2	3	4
2	3	2	2
2	4	4	5
3	5	5	3
3	6	6	6

2. Do Initial Contacts Matter for Gender Peer Effects in Higher Education?

Petra Thiemann

Abstract

This paper examines peer effects in academic outcomes resulting from randomly assigned freshmen week groups at a university. The analysis is based on a dataset of 6 cohorts of students at the University of St. Gallen (CH). Students spend a substantial part of their first week in teams of on average 15 individuals. The analysis exploits random variation in group composition in order to infer the effects of group segregation with respect to gender. Furthermore, the paper investigates whether beneficial group reallocations exist, based on the nonparametric framework by Graham, Imbens and Ridder (2010). The results suggest that female students benefit from higher shares of female peers, in particular with respect to their probability of passing the first year, and with respect to their math grade (results are significant at the 10% level). By contrast, male students are unaffected by changes in the share of female peers. As a result of local increases in segregation with respect to gender, I find small positive effects on grades for subgroups of students, and small positive effects on equality, but the results suffer from imprecision. Thus, no clear assignment rule can be established.

JEL codes: I21, I23, J24
Keywords: segregation, inequality, gender, peer effects, higher education

2.1 Introduction

The topic of peer effects in education has been present both in the scientific literature and in the public debate for many decades. The greatest part of the literature deals with class or school composition according to individual characteristics such as ability (Lavy et al., 2008), gender (Whitmore, 2005) or ethnicity (Angrist and Lang, 2004, Hoxby, 2000). Whereas a large body of literature on peer effects in the area of primary or secondary education exists, more recently the literature has also focused on peer effects in higher education (Epple and Romano, 2011, Sacerdote, 2011). This paper builds on this literature, using a natural experiment at a Swiss university.

The reason why peer effects are an important topic in the economics of education literature is that peers are considered an input into the education production function and are thus crucial for the formation of human capital (Lazear, 2001). Policy makers should therefore be interested in the importance of peer spillovers, also compared to other inputs such as infrastructure and teachers. Depending on the nature of the peer effect, different policies may, or may not be socially desirable. On the one hand, discovering that peer effects are positive in certain settings, policy makers might want to increase the mere amount of peer spillovers by encouraging social interactions. On the other hand, if peer effects depend more on group mixtures than on the intensity of interactions, the externalities created by peers might give rise to policies that change group composition by increasing or decreasing segregation (henceforth “reallocation” policies). Segregation refers to the composition of groups according to specific characteristics. For example, a population is maximally segregated according to gender if the respective groups within the population (e.g., classrooms) consist of only males or only females.

This paper analyzes the impact of gender group composition during the freshmen week at the University of St. Gallen on academic outcomes (i.e., grades, retention, major choice). In particular, the paper examines whether adjusting the share of female students in a peer group can improve a female student’s academic performance. Female students are underrepresented among university entrants in St. Gallen; on average, only 32% of all freshmen are female. Furthermore, only 61% of female students pass the first year; this fraction is 7 percentage points lower than the passing rate among male students. This pattern can neither be explained by differences in socio-demographic characteristics, nor by differences in high school GPA. A growing literature on the role of homophily for friendship formation (Currarini et al., 2009, Carrell et al., 2013) and a well-established literature on the role of social ties for student retention (Tinto, 1975) motivate the hypothesis that this pattern partly arises because female students have more difficulties to find friends and study partners in a primarily male environment.

Peer effects are difficult to identify because individuals typically select into their peer groups. To solve the identification problem, this paper exploits a natural experiment at the University of St. Gallen. Freshmen are randomly assigned to groups of on average 15 students, in which they spend most of the week. This institutional setup will be used in order to implement two types of analyses. First, random assignment to peer groups allows for the study of peer effects in a regression model. I regress educational outcomes on variables that capture group composition with respect to gender. Second, I apply the nonparametric framework of Graham et al. (2010) in order to study reallocation effects. The database consists of administrative records from 6 cohorts of freshmen, that is, 5,012 students in total.

This paper contributes to the literature in two ways. First, the study examines at a relatively short intervention, compared to other studies where interactions occur over a period of one year or more. The one-week intervention studied here exposes students to a certain set of other students when they enter the university, but students may still choose friends and study partners among the full entering cohort. Other policies exploited before, especially studies on military cohort composition, are more restrictive and thus more invasive (Carrell et al., 2009, Lyle, 2009). Second, this study contributes to the discussion on gender segregation effects in higher education. So far, only one study examines gender peer effects in higher education (Oosterbeek and van Ewijk, 2014), but the authors do not explicitly focus on segregation effects. In particular, this is the first study to apply the nonparametric framework outlined in Graham et al. (2010) to a higher education setting. The advantage of the nonparametric method applied here is that no a priori functional form assumptions are necessary. This is important in a context where no clear theoretical predictions on peer and segregation effects exist.

The results suggest that small gender peer effects exist. Female students in groups with higher shares of females perform overall slightly better, with the largest effects for math grades and for the probability of passing the first year. The results are significant at the 10%-level. Male students are not affected by the share of female peers. Consistent with the results on peer effects, the segregation analysis shows that increasing segregation toward more gender separated groups might increase course grades during the first semester for some students, and might reduce overall inequality in grades between females and males (significant at the 10%-level). Yet, due to the rather high imprecision of the results, this study does not derive any policy conclusions from this finding.

The paper is structured as follows. After discussing the existing literature in Section 2.2, Section 2.3 explains the institutional setup. The paper proceeds by a presentation of the data and descriptive statistics in Section 2.4, followed by an explanation of the identification and estimation strategy in Section 2.5. Sections 2.6 and 2.7 present and discuss the results.

2.2 Literature

The literature on peer effects in higher education has grown substantially during the last ten years. It is distinct from the literature in primary or secondary education for two reasons. First, whereas peers are most often associated with classmates in the primary/secondary education literature, the definition of peers is more divers and less obvious in the case of higher education. Peers can be roommates (Sacerdote, 2001), members of squadrons in a military academy (Carrell et al., 2009) or members of cohorts in medical schools (Arcidiacono and Nicholson, 2005), members of the same working group (Oosterbeek and van Ewijk, 2014), as well as individuals attending the same lectures (De Giorgi et al., 2010). In either case, the definition of peers depends on the institutional setting, and different institutional settings lead to different kinds of treatments and identification strategies. In the following, I discuss the literature on peer effects in higher education more deeply, with a particular focus on the methodology used and on the question whether segregation effects have been looked at so far. Segregation effects are of particular interest as individuals might sort themselves to a group or be assigned according to whether their peers have similar characteristics (see, for example, Graham (2011) for a discussion of "assortative matching"). Policy interventions might then promote or countervail the sorting or matching of individuals. Starting from a summary of the methodology, I explore the results, advantages and drawbacks of existing studies in the field.

2.2.1 Methodology: Identification of peer effects and the effects of segregation

From the methodological point of view, most of the literature uses linear-in-means type models to identify marginal effects of changes in peer composition (Manski, 1993). Yet, imposing a purely linear relationship between exogenous peer characteristics and individual outcomes means excluding the existence of positive or negative net effects of reallocations between peer groups. If, for example, the share of high ability students were increased in one group, the share of high ability students decreases in all other groups by the same amount, given same size groups. A linear model would predict a zero net effect of the policy (Hoxby, 2000). As emphasized in Graham et al. (2010), purely linear models would thus miss one of the most important points policy makers are interested in: the (marginal) effects of segregation.

Research on peer effects has been aware of the importance of segregation effects, but has for a long time been lacking the methodological instruments to systematically identify and estimate segregation effects. The route taken by most researchers is to approach segregation effects via effect heterogeneity (Sacerdote, 2001, Zimmerman, 2003, Carrell et al., 2009). If one observes that, for example, women benefit more from having a higher fraction of women in the group than boys suffer from having a lower fraction of women, increasing segregation might induce a positive effect. Effect heterogeneity can for example be measured by splitting up the sample according to the stratifying variable and by then running a regression within the two subsamples. Notice that, in order for effect heterogeneity to be informative with respect to segregation effects, the characteristics of the stratifying variable (e.g. gender) have to coincide with the characteristics of the treatment variable (e.g. share of females in a group, gender of the roommate).

Graham (2011) shows how to compute segregation effects from regression estimates under this type of stratification. He extends his results also to the case of a non-linear specification of the treatment variable (e.g. as indicator variable, in quadratic or cubic terms). Nonparametric identification of segregation effects is even more desirable, but has not been applied to the study of peer effects in higher education so far. The reason might be that, on the one hand, the nonparametric identification framework for segregation effects as presented in Graham et al. (2010) has only recently been available. On the other hand, the identifying assumptions are very specific and not easy to satisfy. In particular, the framework relies on random variation in peer group shares with respect to binary individual characteristics (e.g. gender). In order to identify the effects nonparametrically, continuous variation across the support of the shares is required. Therefore, this approach is unsuitable for very small peer groups, and has therefore not been applied to roommate studies and similar small-group settings.

2.2.2 Empirical studies

Random assignment to roommates has frequently been exploited in the peer effects literature. Investigations of bigger groups as in the setting presented here are less common.

As argued in Stinebrickner and Stinebrickner (2006), students tend to spend a lot of time with their roommates, but roommates might still not constitute a relevant peer group, because interactions might arise only out of necessity. Studies looking at other peer-group definitions therefore complement the results from roommate studies. Carrell et al. (2009) examine peer effects in squadrons in the US Airforce Academy. In line with the findings of the seminal roommate study by Sacerdote (2001), the authors find positive ability peer effects, with ability being measured by verbal SAT scores. Moreover, Carrell et al. (2009) report effect heterogeneity with respect to ability, but not with respect to gender or ethnicity. Conducting a study in a similar environment, the US military academy at Westpoint, Lyle (2007) neither supports the existence of peer effects nor the existence of effect heterogeneity.

In contrast to the findings by Carrell et al. (2009) as well as Lyle (2007), two studies support the existence of effect heterogeneity with respect to gender. Oosterbeek and van Ewijk (2014) exploit randomization of the share of female students in work groups at the University of Amsterdam. They find that only boys' math performance decreases in the share of girls, but girls' performance remains constant. Moreover, for classes in cohorts of US medical schools, Arcidiacono and Nicholson (2005) find peer effects only for female students, using a fixed-effects model.

None of the studies cited so far systematically assesses the effects of reallocations, with the exception of a unique study by Carrell et al. (2013). The authors present the results of a field experiment at the US Airforce Academy where peer groups are optimally assigned to improve the weakest students' educational outcomes. The assignment is based on results of previous regression analyses as discussed by Carrell et al. (2009). Yet, the optimal assignment does not lead to better outcomes for weak students. Carrell et al. (2013) conclude that within-group sorting into peer groups might be one reason for this finding. They further argue that the models used in their prior analyses were not able to capture within-group sorting, and were thus not suitable for extrapolation of the findings to new types of groups. In particular, the new groups created in the reallocation experiment resembled groups in prior cohorts with respect to group means of SAT scores, but not with respect to within-group distributions of SAT scores.

Because of the extrapolation problem, this paper only studies marginal reallocations, that is, reallocations that marginally change the degree of segregation. In other words, this paper only derives results in a local environment around the status quo, following Graham et al. (2010). The results might thus induce a policy maker to make incremental changes to the allocation. The analysis, however, remains silent on the issue of global reallocations that strongly deviate from the status quo.

2.3 Institutional setup

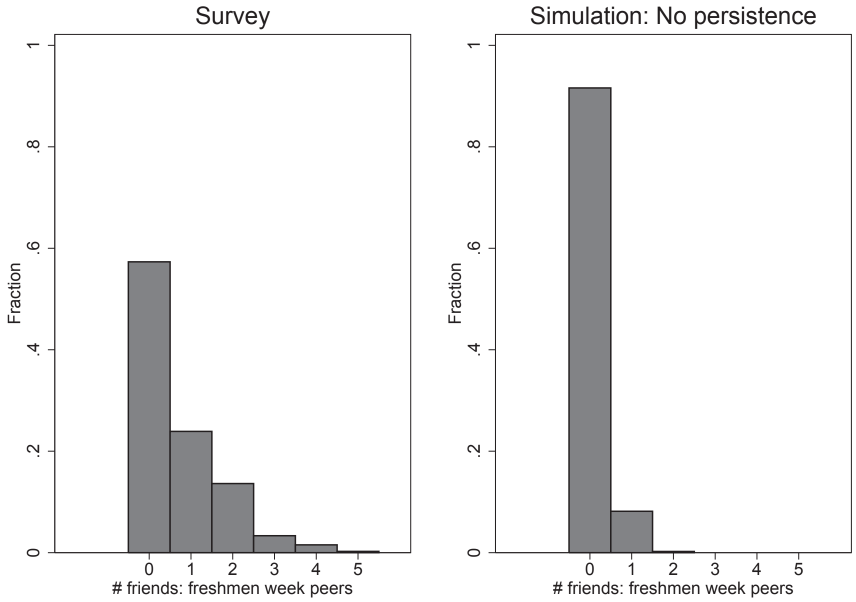
The University of St. Gallen offers undergraduate and graduate studies in Business Administration, Economics, International Affairs, Law and Economics, as well as Legal Studies. The number of newly enrolled undergraduate students is steadily increasing, reaching over 1,000 students in 2009. Since 2001, undergraduate studies start with a mandatory freshmen week. This week aims at introducing the students to the infrastructure (e.g. library, online tools), giving them the opportunity to get to know their fellow students in small groups, and introducing them to group work, which makes up an important part of the studies at the University of St. Gallen. At the beginning of the week, individuals are assigned to small teams of on average 15 individuals in which they perform many of the freshmen week activities. The number of teams amounts to approximately 60 per year. The assignment is quasi-random, that is, based on last names, and conditional on gender and admission rule (i.e. whether students had to complete an entrance test to be admitted, which is the case for non-Swiss students). Details on the randomization procedure can be found in the appendix (see Figure 2.A.1).

All teams perform the same set of activities throughout the weeks. Schedules for the different teams might differ as not all groups can perform the same activities at the same time for logistic reasons (e.g., a library introduction), but the amount of time assigned to the different activities does not vary between teams. The whole week centers around an incentivized case study competition between teams.

Overall, the freshmen week gives rise to intense interaction within teams. The organized activities during the week amount to 57 hours, including evening events. 62% of the program is designated to team activities. Participation in the freshmen week is mandatory, and students are informed in advance by mail that they have to reserve the full week. Two tutors, usually advanced Bachelor or Master students, supervise each group and make sure that students comply to the group activities.

After the freshmen week, students enter their first year, the assessment year. During this year, all students except law students complete the same courses. Students who successfully pass the first year enter their second year and can complete the whole Bachelor degree within at least two more years. Major choice takes place during the last weeks of the assessment year.

Figure 2.1: Persistence of peer groups: survey evidence



The panels shows the distribution of the number of freshmen team members among a students' five best friends. Left panel: Survey results, cross section of Bachelor students, n = 388 (response rate: 19%). Right panel: Simulation with administrative records under the assumption that freshmen groups play no role for the formation of friendships, 6 cohorts (2003, 2004, 2006–2009), n = 5,012. Source: Own calculation using survey and administrative data from the University of St. Gallen.

How important are freshmen groups throughout the further university career? To answer this question, I conducted an online survey among a cross-section of Bachelor students at the University of St. Gallen in May 2012. 65% of survey respondents report that they still engage in free-time activities with at least one member of their freshmen week group; this fraction is substantial, given that the survey was carried out up to three years after the end of the freshmen week. Most importantly, freshmen week members are over-represented among friends (see Figure 2.3). To derive this result, I simulate a random allocation of friendships within a cohort of students and compare this random allocation to the actual distribution of friendships, as indicated by the survey results. I concentrate on the five best friends. If freshmen group members had no impact on the formation of study partnerships, only about 10% of students would be friends with at least one freshmen group member. By contrast, a much larger fraction, that is, more than 40% of survey participants, indicate that a least one of their five best friends comes from their freshmen week group. While these results are non-representative and therefore have to be interpreted with caution, they still suggest a long-lasting impact of the freshmen week.

One limitation to the definition of peer groups in this papers is that that freshmen groups do not fully coincide with actual friendships. Unfortunately, friendship networks are unobserved in our data. Yet, the analysis is relevant from a policy perspective as university administrations can manipulate team arrangements during the first week, but cannot directly influence friendship networks.

2.4 Data and descriptive statistics

2.4.1 Dataset

The analysis is based on administrative data records from the University of St. Gallen, combined with information on freshmen team assignment. The dataset contains detailed information on semester enrollment, course choice, academic performance, that is, grades in all courses, as well as background characteristics of the students. The dataset includes the full population of entering freshmen of the cohorts of 2003, 2004, and 2006-2009. Information on freshmen group assignment is unfortunately missing for the years 2001, 2002 and 2005, and the full set of outcomes is only observed up to 2009.

Table 2.A.1 shows the fraction of entering students per year that are included in the estimation sample. Only three criteria lead to exclusion from the analysis: First, I exclude students that could not be matched to the freshmen group file. Second, I delete special groups that are selected in advance. These groups are, for example, groups for students that serve in the military and therefore do not participate fully in the freshmen week. Third, I exclude one group with a fraction of male students of 100%, as this group is an outlier in terms of the treatment. All criteria combined lead to an exclusion of 4% of students from the estimation sample. The sample size amounts to 5,012 individuals in 345 groups. The average group size is 15, with a minimum of 7 and a maximum of 22 group members.

2.4.2 Background characteristics

The administrative dataset contains the following socio-demographic background characteristics: gender, age, nationality (Swiss, German/Austrian, other), mother tongue (German, other), place at which the university entrance degree (e.g., high school diploma) has been completed (for Swiss students: canton is reported, for all others: abroad), and whether students had to complete an entrance exam to be admitted. According to the university regulations, students with both a foreign entrance degree and a foreign nationality have to pass an entrance exam. All other students are automatically admitted due to cantonal legislation. Furthermore, I observe whether students have decided in advance to complete the "extended track", that is, to complete all first-year courses within two years. This option is only open to students with a non-German mother tongue, as first-year courses are held in German.

Table 2.A.2 shows that average background characteristics as well as average outcomes stay relatively stable across cohorts, which justifies the approach to use a pooled cross section. Female students are under-represented in all years (32% of all freshmen on average). I observe small time variation for some variables: The extended track has become more popular over time. Moreover, group size has grown over time, owing to growing cohort sizes. To control for cohort effects, I include cohort dummies throughout the further analysis.

2.4.3 Outcomes

From the enrollment and course files, I infer four sets of outcomes: First, I define four binary outcomes that capture student retention during the first year, based on course records during the first year. The first year consists of two semesters. During these two semesters, students can drop out voluntarily, or exceed a threshold value for failed courses. I define (i) a variable “first semester completed”, which indicates whether a student completed all required courses during the first semester, (ii) a variable “‘first semester passed”, which indicates whether a student has completed all courses, and did not exceed the threshold value for failed courses, (iii) a variable “second semester passed”, which indicates that a student has passed the first semester and completed all second semester courses, and (iv) a variable “second semester passed”, which indicates that a students has been promoted to the second year in their first attempt. Students who do not pass the first year in the first attempt can repeat the first year.

Second, I define a set of binary variables on major choice, which takes place at the end of the first year. Students can choose between Business, Economics, International Affairs, Law and Economics, and Legal Studies. The most popular majors are Business (52% of entrants enroll for this major), Economics (12%), and International Affairs (11%); Law and Economics as well as Legal Studies are less popular (5% each). To achieve sufficient variation in the outcome, I therefore concentrate on the three most popular majors throughout the further analysis. Moreover, I assess whether students enter the second year at all. The fraction of students who enter the second year exceeds the fraction of students who pass the first year in the first attempt (66% versus 79%) because of the option to repeat.

Third, I assess course grades during the first semester as immediate performance predictors. I focus on the four core courses Math, Economics, Business, and Legal Studies. Notice that course grades are only defined for individuals who complete the respective exam. As most students complete these courses (92% complete Math, 98% complete Economics, 99% complete Business, and 97% complete Legal Studies), I assume that the selection bias arising from this definition will be rather small. I standardize course grades at the cohort-course level.

Fourth, to capture performance effects over a longer time horizon, I analyze average grades (GPA) during all semesters of the first and the second year. Grades for the first and second semester are defined for all individuals who have taken at least one exam (99%). Grades are weighted by credit points and standardized at the cohort level. Second year grades are defined for all students who have taken an exam during the respective semester. Because of high dropout rates, these outcomes are defined only for 75% and 77% of students, respectively. Results for these outcomes might therefore suffer from selection bias. All GPAs are standardized at the grade-cohort level.

Table 2.1: Descriptive statistics and tests for random assignment – female students

	(1)	(2)	(3)	(4)	(5)
	Mean	Share female		Diff.	t-stat.
		<= 0.3	> 0.3	(adj.)	
Background characteristics					
Age	20.01	20.10	19.86	-0.14	-1.17
Entrance exam	0.12	0.11	0.13	-	-
Non-Swiss nationality	0.19	0.17	0.22	0.03	2.31
Non-German mothertongue	0.12	0.12	0.12	-0.01	-0.39
High school St. Gallen	0.17	0.17	0.17	0.01	0.30
Extended track	0.06	0.06	0.06	-0.01	-0.71
Group variables					
Share female	0.28	0.24	0.35	0.11	20.03
Group size	15.32	14.85	16.04	0.22	1.17
Outcomes					
1st semester completed	0.92	0.91	0.93	0.02	1.17
1st semester passed	0.79	0.78	0.79	0.01	0.51
2nd semester completed	0.70	0.69	0.71	0.03	1.29
2nd semester passed	0.61	0.59	0.63	0.04	1.61
Enters 2nd year	0.73	0.72	0.74	0.02	1.09
Business	0.43	0.43	0.43	0.00	-0.03
Economics	0.09	0.08	0.09	0.01	0.63
International Affairs	0.14	0.12	0.16	0.03	1.89
Law & Econ.	0.05	0.05	0.05	0.00	0.21
Legal Studies	0.07	0.08	0.07	0.00	-0.31
Grade math	-0.04	-0.08	0.02	0.09	1.74
Grade economics	-0.23	-0.24	-0.20	0.01	0.16
Grade business	-0.06	-0.08	-0.03	0.03	0.67
Grade legal studies	-0.05	-0.08	-0.02	0.03	0.63

Continued on the next page.

Continued from the previous page: Descriptive statistics and tests for random assignment – female students

	(1)	(2)	(3)	(4)	(5)
	Mean	Share female		Diff.	t-stat.
		<= 0.3	> 0.3	(adj.)	
Background characteristics					
Outcomes (continued)					
GPA semester 1	-0.10	-0.13	-0.06	0.05	1.00
GPA semester 2	-0.09	-0.12	-0.05	0.05	0.95
GPA semester 3	-0.02	0.00	-0.04	-0.07	-1.24
GPA semester 4	-0.02	-0.02	-0.02	-0.02	-0.31

The table shows mean student characteristic, group variables, and outcome variables by cohort (year of entry) for the sample of female students (1,581 observations). "Share female" is the treatment variable and computed as the leave-own-out mean. Course grades are standardized at the course level. Grades are missing for dropouts and for individuals who did not take the respective exam. Average grades (GPA) are defined for all individuals who take at least one exam in the respective academic years and are standardized at the cohort level. Column 2 (column 3) shows means of variables for females in groups with a share of female peers below 30% (above 30%), corresponding to 957 observations (624 observations). Column 4 shows the differences in means, which are adjusted for the stratifying variable ("entrance exam") as well as for cohort effects. I.e., the differences are average partial effects from a linear regression of the background variables on the binary treatment variable "share > 0.3". I fully interact the treatment variable with the variable "entrance exam" and linearly control for cohort dummies. Column 5 reports the t-statistics of the coefficient reported in column 4, indicating whether the adjusted difference is statistically different from zero. The t-statistic is based on standard errors that are clustered at the group level. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.2: Descriptive statistics and tests for random assignment – male students

	(1)	(2)	(3)	(4)	(5)
	Mean	Share female		Diff.	t-stat.
		<= 0.3	> 0.3	(adj.)	
Background characteristics					
Age	20.29	20.29	20.30	0.05	0.88
Entrance exam	0.21	0.22	0.21	-	-
Non-Swiss nationality	0.27	0.28	0.26	-0.01	-1.41
Non-German mothertongue	0.10	0.10	0.10	0.00	-0.39
High school St. Gallen	0.15	0.15	0.14	-0.01	-0.75
Extended track	0.05	0.06	0.05	-0.01	-1.53
Group variables					
Share female	0.33	0.25	0.36	0.11	19.81
Group size	15.15	14.70	15.33	0.35	2.01
Outcomes					
1st semester completed	0.93	0.94	0.93	-0.01	-0.70
1st semester passed	0.83	0.84	0.83	-0.02	-1.01
2nd semester completed	0.77	0.78	0.77	-0.01	-0.81
2nd semester passed	0.68	0.69	0.68	-0.02	-1.03
Bachelor entry	0.81	0.81	0.81	-0.02	-1.20
Business	0.56	0.55	0.57	0.00	-0.13
Economics	0.14	0.14	0.14	0.01	0.78
International Affairs	0.09	0.09	0.09	0.00	-0.29
Law & Econ.	0.05	0.05	0.05	-0.01	-0.62
Legal Studies	0.04	0.05	0.03	-0.01	-1.76
Grade math	0.02	0.00	0.03	0.04	0.87
Grade economics	0.11	0.12	0.11	0.00	0.11
Grade business	0.04	0.04	0.04	0.00	0.04
Grade legal studies	0.03	0.04	0.03	-0.01	-0.29

Continued on the next page.

Continued from the previous page: Descriptive statistics and tests for random assignment – male students

	(1)	(2)	(3)	(4)	(5)
	Mean	Share female		Diff.	t-stat.
		<= 0.3	> 0.3	(adj.)	
Background characteristics					
Outcomes					
GPA semester 1	0.06	0.06	0.06	0.01	0.20
GPA semester 2	0.05	0.05	0.05	0.00	0.04
GPA semester 3	0.01	0.00	0.01	0.02	0.44
GPA semester 4	0.01	0.01	0.00	-0.05	-0.97

The table shows mean student characteristic, group variables, and outcome variables by cohort (year of entry) for the sample of male students (3,431 observations). "Share female" is the treatment variable and computed as the leave-own-out mean. Course grades are standardized at the course level. Grades are missing for dropouts and for individuals who did not take the respective exam. Average grades (GPA) are defined for all individuals who take at least one exam in the respective academic years and standardized at the cohort level. Column 2 (column 3) shows means of variables for males in groups with a share of female peers below 30% (above 30%), corresponding to 986 observations (2,445 observations). Column 4 shows the differences in means, which are adjusted for the stratifying variable ("entrance test") as well as for cohort effects. I.e., the differences are average partial effects from a linear regression of the background variables on the binary treatment variable "share > 0.3". I fully interact the treatment variable with the variable "entrance test" and linearly control for cohort dummies. Column 5 reports the t-statistics of the coefficient reported in column 4, indicating whether the adjusted difference is statistically different from zero. The t-statistic is based on standard errors that are clustered at the group level. Source: Own calculations using administrative data from the University of St. Gallen.

2.4.4 The gender gap in outcomes

A comparison of Tables 2.1 and 2.2 depicts large differences in outcomes between male and female students. For example, females students are 8 percentage points less likely to pass the first year in their first attempt and 8 percentage points less likely to enter the second year. Moreover, they perform worse in all analyzed first-year courses, and also obtain a lower GPA throughout the first and the second year. Only in the second year, the gap closes almost entirely, probably because lower-performing female students have dropped out.

A natural explanation for this pattern might lie in differences in observable and unobservable characteristics, but the characteristics available in the data cannot explain the strong differences: While the raw difference in the probability of passing the first year amounts to 8 percentage points, the difference shrinks only by 2 percentage points when controlling for all characteristics available in the administrative data (Table 2.A.3). As high school grades are unavailable in the administrative data, Table 2.A.4 presents an additional analysis for one recent cohort (2011), for which we collected data on high school GPA. Controlling for high school GPA, the gap between male and female students even widens, as female students perform on average better in high school than male students. From these observations, two hypotheses emerge: Either male and female students differ in unobservable characteristics that create a performance advantage for males; or female students find it more difficult to find like-minded friends and study partners, as they represent the minority of students. The question whether a higher share of female peers could help female students to perform better remains ultimately an empirical question.

2.4.5 Treatment

To study the response of students' performance to gender group composition, I define a treatment variable, that is, the share of female students in a student's freshmen group, computed as the leave-own-out mean. This share varies between 0.07 and 0.57, with an overall mean of 0.32 and a standard deviation of 0.07 (Figure 2.A.2). Because of the stratification, the mean for female students is slightly lower than the mean for male students (0.28 versus 0.33, see Tables 2.1 and 2.2). Throughout the further analysis, I include the treatment both as a continuous variable, and as a dummy variable, defined as high share (above 0.30) and low share (below 0.30). The cut-off of 0.30 is convenient, as it is a round number close to both the mean (0.32) and the median (0.31), and at the same time identical to the cut-off chosen in a similar study by Oosterbeek and van Ewijk (2014). On average, moving from a group with a low share of females to a group with a high share of females creates a difference in the share of females by 0.11 (see Tables 2.1 and 2.2).

As discussed in Section 2.3, the assignment to groups is quasi-random, conditional on gender and on the admission rule (variable “entrance exam”). To test the random assignment assumption, Tables 2.1 and 2.2 assess whether background variables are balanced across individuals in groups with low and high shares of female peers, respectively, conditional on the admission rule. Except for the nationality, all background variables balance well. To adjust for potential biases due to this imbalance, I therefore control for nationality throughout the further analysis, in addition to controlling for the admission rule and for year dummies.

A comparison of outcome variables across groups with high and low levels of the treatment reveals some differences that are significant at the 10%-level: Female students with a higher share of females are more likely to choose an International Affairs major, which is more popular among women, and they perform better in math. By contrast, male students are only slightly less likely to major in Legal Studies if assigned to groups with higher shares of females. Overall, Tables 2.1 and 2.2 suggest that differences in outcomes across groups with different shares of females are rather modest.

2.5 Identification strategy and estimation

2.5.1 Identification

The first part of the analysis is based on a simple model which is derived from the linear-in-means model (Manski, 1993). The first model assumes that the outcome Y_{ig} of individual i in group g depends on D_{ig}, which is a binary indicator for a share of female students above 0.3 in an individual's group, on a vector of individual control variables x_i, and on a vector of cohort dummies C_i. ϵ_{ig} is an ideosyncratic error term:

$$Y_{ig} = \gamma D_{ig} + x_i'\beta + C_i'\alpha + \epsilon_{ig}, \tag{2.1}$$

with $D_{ig} = \mathbf{1}(\text{share\_female}_{ig} > 0.3)$, where share_female_{ig} is the share of females in a group, and $\mathbf{1}(.)$ is the indicator function.

This model is valid for continuous outcomes. For binary outcomes, the model may be changed to:

$$Y_{ig} = \mathbf{1}\left(\gamma D_{ig} + x_i'\beta + C_i'\alpha - \epsilon_{ig} > 0\right). \tag{2.2}$$

Transforming a continuous treatment variable into a dummy variable is common across similar studies (Carrell et al., 2009, Oosterbeek and van Ewijk, 2014), and facilitates an easy interpretation of the treatment effect. If assignment to peer groups is (conditionally) random, and if shocks to the outcome are ideosyncratic shocks unrelated to the share of females, the coefficient γ identifies an "exogenous" peer effect, that is, an effect of peer group composition in terms of observable group characteristics, as established in earlier papers on this topic (Manski, 1993, Sacerdote, 2011). The analysis is conducted for males and females separately; thus, the analysis can capture effect heterogeneity and can be informative for reallocation effects: Only if responses for males and females differ, reallocations of the shares of females across groups can be welfare improving (see Section 2.2).

I also test a second, alternative model that captures responses to incremental changes in the share of females, rather than responses to a "jump" from a low to a high share. This model includes the share of females as a continuous variable and models the response function as a kth-order polynomial of this variable:

$$Y_{ig} = \sum_{k=1}^{K} \gamma_k \text{share\_female}_{ig}^{k} + x_i'\beta + C_i'\alpha + \epsilon_{ig}. \tag{2.3}$$

Similar to the analysis presented above, the vector γ_k determines the shape of the response function to the share of females and thus captures exogenous peer effects. A corresponding model for binary outcomes can be written as:

$$Y_{ig} = \mathbf{1}\left(\sum_{k=1}^{K} \gamma_k \text{share\_female}_{ig}^{k} + x_i'\beta + C_i'\alpha - \epsilon_{ig} > 0\right). \tag{2.4}$$

As the shape of the response function is unknown (to be estimated), I specify different values for K, in particular, $K \in \{1, ..., 3\}$. Notice that the choice of the polynomial order, K, is subject to a familiar trade-off: A higher polynomial order might lead to less biased estimates, but also to lower precision. Due to limited sample sizes, I restrict the analysis to at most a third-order polynomial.

The second part of the analysis, that is, the analysis of segregation effects, is based on the identification framework by Graham et al. (2010). The authors show that non-parametric identification of peer and segregation effects can be achieved under six assumptions: (1) *no cross neighborhood spillovers*, (2) *within-type peer exchangeability*, (3) *inclusive definition of type*, (4) *no matching and sorting on unobservables*, (5) *continuous variation*, and (6) *random sampling.* I will argue in the following that these assumptions hold in this setting. In order to simplify the argument, I follow the authors' terminology of "types" and denote female students as type-F-students and male students as type-M-students.

The assumption of *no cross neighborhood spillovers* means that peer spillovers within a group potentially exist, but spillovers between groups do not. To put it more precisely, group composition might affect individual outcomes, but feasible reallocations of individuals between the other groups must not affect individual outcomes. A reallocation is defined as feasible if the overall distribution of types across all groups remains constant, compared to the initial distribution. This assumption implies that the outcomes of freshmen group members must not change if, for example, the ability composition within other groups changes, holding the overall ability distribution across groups fixed. I argue that this assumption holds, as the groups never directly interact with other groups. Individuals stay either together in their groups, or they interact during plenary sessions or events in which every freshman participates.

Within-type peer exchangeability denotes that peers of the same type are a priori equally influential with respect to individual outcomes. This assumption implies that peer groups have a priori no hierarchical structure with differential roles of peers. Specifically, the assumption requires that peers are independently and identically distributed within groups, given their type and possibly any background characteristics. For example, the assumption excludes that the order in which peers are assigned to a group matters. The assumption is likely to hold in our setting as the institutional setting imposes no structure on the peer group when assigning individuals to the groups.

The third assumption requires that the definition of types must be inclusive in the sense that the definition of types captures all unobserved characteristics that are related to that type. To illustrate this point with an example from Graham et al. (2010), suppose that males are more disruptive than females, but disruptiveness is unobserved. Then, the extra-disruptiveness that is introduced in a class by an increase in the share of males is included in the treatment definition. This assumption is important as, together with the assumption of no matching and no sorting, it ensures independence of the treatment from unobserved background characteristics, and thus leads to a proper ceteris-paribus interpretation of the treatment effect. As Graham et al. (2010) point out, it is furthermore unrestrictive in the sense that the assumption applies without loss of generality. In the setting studied here, the assumption holds only conditional on the stratifying variable, "entrance exam".

The assumption of *no matching and sorting on unobservables* refers to the selection of individuals into groups. No matching ensures that individuals do no self-select or are selected into groups based on their type and or on unobservable characteristics related to their type. Matching takes place if individuals select their group according to unobserved group characteristics such as rooms or the teachers, whereas sorting occurs if individuals select their group according to unobserved characteristics of the group members. Both cases can be excluded in our setting, conditional on the variable “entrance exam”. Students are conditionally randomly assigned to groups and have to comply to the groups. Therefore, *no sorting* holds due to conditional random assignment. Moreover, group tutors are randomly assigned as well, so that the no-matching-assumption is equally plausible.

Furthermore, the *continuous-variation*-assumption has to be satisfied. The variation of the fraction of types has to be approximately continuous on the support of the distribution of type fractions across groups. Figure 2.A.2 shows in how far the *continuous-variation* assumption is satisfied. The assumption is valid when aggregating over all years. Then, the support for the share of male students lies between a share of 7% and 47%.

The *random sampling* assumption applies only in so far as the sample used here can be interpreted as a subpopulation from a meta-population. Since I use nearly the full sample of freshmen in the respective cohorts, the interpretation of this assumption has to be clarified. We could argue that the cohorts we look at are drawn from a larger student pool that, for example, also includes potential applicants. Standard errors then reflect uncertainty in the selection of students from the universe of potential students.

All identification results presented in Graham et al. (2010) are based on the idea that each individual’s outcome can be represented as a function of the allocation of individuals to groups. This allocation response function relates the outcome of an individual to the features of a specific group allocation, i.e. the types and unobserved characteristics of his peers, the types and unobserved characteristics as well as the group allocation of all other individuals, the features of all groups (e.g. teacher quality), and unobserved characteristics of the individual. If the identifying assumptions were holding unconditionally, the (possibly non-linear) type-specific “mean allocation response functions”, that is, the mean response for individuals of a specific type to the allocation given is identified as

$$m_F(s) = E[Y_i|T_i = F, S_i = s], \tag{2.5}$$

or

$$m_M(s) = E[Y_i|T_i = M, S_i = s], \tag{2.6}$$

respectively, where s denotes the share of female peers, T_i denotes the individual's type, and Y_i denotes the individual outcome. Types are denoted by either M or F, so that $T \in M, F$, which allows the functions to potentially differ by type. See Graham et al. (2010) for a detailed derivation of the identification results.

This identification result is strong in the sense that the mean allocation response function is identified as long as one knows each individual's type and group. In the case that the assumptions of *no matching and no sorting* and *inclusive definition of type* hold only conditionally on observed characteristics (individual characteristics W_i, or group characteristics X_i), the identification results carry through when conditioning on the respective covariates, so that

$$m_F(s) = E_{\underline{W},X}[Y_i|T_i = F, S_i = s, \underline{W}_i, X_i], \tag{2.7}$$

where vector $\underline{W}_i$ denotes the characteristics of all members of a peer group of individual i. The result for $m_M(s)$ follows analogously (Graham et al., 2010).

Identification of the mean allocation response functions allows for the identification of different measures of allocation efficiency. All measures presented in the paper are valid only within the bounds of the continuous support of the type distribution. First, Graham et al. (2010) present a measure of average spillover strength (ASE), which is an alternative to average partial peer effects from models 2.1-2.4. The average spillover strength captures the externalities that are induced by marginally changing group composition in one direction (e.g. marginally increasing the share of males in all groups). The average spillover effect is thus given as:

$$\beta^{ase} = E[d_k(S_i)\{S_i \nabla_s m_M(S_i) + (1 - S_i)\nabla_s m_F(S_i)\}], \tag{2.8}$$

where $\nabla_s m_M(S_i)$ denotes the first derivative of the mean allocation response function with respect to s, and $d_k(S_i)$ is an indicator function that ensures that observations close to the boundary or off the support of S are not included in the analysis. This trimming function ensures sensible estimation results. Yet, it has to be included in the identification result as well in order to point out that the result holds only within the continuous support of S. Identification for regions off the support is infeasible.

Notice that the average spillover effect presented here does not correspond to a feasible reallocation of individuals. This drawback is taken into account by the local segregation outcome effect (β^{lsoe}) measuring the effect of a marginal segregation increasing reallocation on average outcomes. The authors decompose the segregation measure in an effect β^{lppe} (local private peer effect) for the individuals who change groups (movers) and a different effect β^{lepe} (local external peer effect) for the individuals who stay in their group (stayers). A third effect under consideration is the local segregation inequality effect (β^{lsie}), measuring how a marginal increase in segregation affects the outcome gap between female and male individuals. I refer to Graham et al. (2010) for the formal definitions of these coefficients.

2.5.2 Estimation

In order to estimate the peer-effects model in Equations 2.1-2.4, I run regressions for male and female students separately. For continuous outcomes, I use OLS estimation, and for binary outcomes, I use probit models. The choice of parametric models is motivated by the use of parametric models in similar studies. Moreover, the parametric models provide more robust results than nonparametric models, even if outcomes suffer from low variation (see Section 2.6). In the parametric estimation, I control linearly for cohort, group size, and nationality, and fully interact the treatment variable with an indicator for the stratifying variable "entrance exam". Standard errors are clustered at the group level and computed using the delta method. The standard errors reflect two types of uncertainty. First, I only observe a subset of students out of a potential universe of students (i.e., I never observe students who do not enroll). Second, the assignment generates uncertainty as well (i.e., for each student, I only observe the actual and never a counterfactual outcome).

The estimation procedure for the segregation model will be based on the procedure suggested by Graham et al. (2010).[34] All estimators for the respective coef-

[34] Matlab code provided by Bryan Graham upon request, version from September 2010.

ficients are shown to be consistent and asymptotically normal in their paper. The estimators of β^{ase}, β^{lsoe} and β^{lsie} are based on the estimators of the mean allocation response functions, $\hat{m}_M(s)$ and $\hat{m}_F(s)$, and their derivatives, $\nabla_s \hat{m}_M(s)$ and $\nabla_s \hat{m}_F(s)$. For example, the estimator $\hat{\beta}^{ase}$ can be written as

$$\hat{\beta}^{ase} = \frac{1}{I} \sum_{i=1}^{I} d_k(S_i) \{ S_i \nabla_s \hat{m}_M(S_i) + (1 - S_i) \nabla_s \hat{m}_F(S_i) \}. \tag{2.9}$$

The mean allocation response functions, $\hat{m}_M(s)$ and $\hat{m}_F(s)$, are estimated nonparametrically, using kernel smoothing methods (see Section 2.B). The bandwidths are determined by leave-own-group-out cross-validation, taking within-group interdependence of individual outcomes into account. Following Graham et al. (2010) I use standard normal kernels. The derivatives $\nabla_s \hat{m}_M(s)$ and $\nabla_s \hat{m}_F(s)$ are not estimated, but instead, the derivatives are computed analytically; all necessary parameters are available from the estimation of $\hat{m}_M(s)$ and $\hat{m}_F(s)$ and plugged into the derivative. The estimation of β^{lsoe} and β^{lsie} proceeds accordingly. Moreover, the authors provide an estimator for the asymptotic variance-covariance matrix. The estimation of this matrix is based on the estimation of an efficient influence function. Standard errors account again for within-group interdependence in outcomes. In all nonparametric estimations, I control for the conditioning variable “entrance exam”. Including further control variables is not possible due to insufficient sample size. For further details on the estimation, see Section 2.B.

2.6 Results

2.6.1 Peer effects

As outlined in Section 2.5, I first assess peer effects separately for females and males. For female students, increasing the share of female peers beyond the threshold of 0.3 has overall positive effects on academic performance, but only three of the average partial effects in Table 2.3 are significant at conventional levels: First, female students with higher shares of female peers are on average 5 percentage points more likely to pass the first year in their first attempt; this increase corresponds to 8%

of the passing rates of female students (61%) and is thus substantial (significant at the 10%-level). Second, female students in groups with higher shares of female peers perform on average 0.09 standard deviations better in math (significant at the 10%-level). Third, female students are more likely to start a major in International Affairs, which is in general a more popular choice among females, compared to males (significant at the 5%-level). The only seemingly negatively affected variables are grades in the second year, probably because of a dropout bias: Lower performing females might be encouraged not to drop out when in groups with more female peers. Yet, the negative effect on grades in the second year is not significant at any conventional level.

In contrast to female students, male students are hardly affected by the share of female peers (Table 2.4). All average partial effects are close to zero, sometimes slightly negative, but never significant at any conventional level.

2.6.2 Peer effects: Alternative models and robustness

How do the average partial effects of the dummy variable analysis compare to average partial effects obtained with alternative models? To compare different modeling approaches, I first assess the average partial effects in models where the share of females is included as a first-, second-, and third-order polynomial. Second, I compare the results of these models to the results of the dummy variable analysis in Section 2.6.1.

The results of these comparisons are as follows: First, the average partial effects are robust to the choice of the polynomial order, for both male and female students, and both in terms of quality and in terms of magnitude (see Tables 2.A.10 and 2.A.11). Furthermore, none of the average partial effects are significant at any conventional significance level, which suggests that the precision of the estimates is not strongly affected either. Second, the model choice with respect to a coding of the treatment as a dummy versus a continuous variable matters in terms of the magnitude, but not in terms of the quality of the effects. In order to compare magnitudes, I divide the average partial effects from the dummy variable model by 0.11, which is the average difference in the share of females. This is a rule-of-thumb approach, but

Table 2.3: Average partial effects of a high share of female peers (coded as dummy variable) on academic performance – female students

(1) Dependent variables: Retention during the first year				
	First semester		Second semester	
	completed	passed	completed	passed
Fraction female > 0.3	0.02	0.01	0.03	0.05*
S.E.	(0.01)	(0.02)	(0.02)	(0.03)
p-value	(0.18)	(0.55)	(0.14)	(0.06)
Obs.	1,581	1,581	1,581	1,581
R-squared	0.04	0.01	0.02	0.04
Controls	Yes	Yes	Yes	Yes

(2) Dependent variables: Major choice				
	Enters 2nd year	Business	Economics	Interational Affairs
Fraction female > 0.3	0.03	0.00	0.01	0.04**
S.E.	(0.02)	(0.03)	(0.01)	(0.02)
p-value	(0.19)	(0.98)	(0.55)	(0.04)
Obs.	1,581	1,581	1,581	1,581
R-squared	0.02	0.02	0.04	0.01
Controls	Yes	Yes	Yes	Yes

(3) Dependent variables: Course grades, first semester				
	Math	Economics	Business	Legal studies
Fraction female > 0.3	0.09*	0.02	0.04	0.04
S.E.	(0.05)	(0.05)	(0.05)	(0.05)
p-value	(0.08)	(0.70)	(0.37)	(0.40)
Obs.	1,375	1,541	1,554	1,517
R-squared	0.05	0.05	0.04	0.04
Controls	Yes	Yes	Yes	Yes

Continued on the next page.

Continued from the previous page: Average partial effects of a high share of female peers (coded as dummy variable) on academic performance – female students

(4) Dependent variables: Average grades, first four semesters				
	1st semester	2nd semester	3rd semester	4th semester
Fraction female > 0.3	0.06	0.06	-0.08	-0.02
S.E.	(0.05)	(0.05)	(0.06)	(0.06)
p-value	(0.21)	(0.23)	(0.18)	(0.75)
Obs.	1,563	1,563	1,102	1,120
R-squared	0.06	0.05	0.06	0.04
Controls	Yes	Yes	Yes	Yes

The table presents average partial effects of an above-average share of female students in the group (binary variable: share of female > 30%) on performance indicators for female students (1,581 observations). The analysis uses probit models for binary performance outcomes (panels 1 and 2) and OLS models for continuous performance outcomes (panels 3 and 4). Grades are standardized at the course-cohort level (panel 3) or cohort level (panel 4). Grades are missing for dropouts and individuals who did not take the respective exam. Average grades are defined for all individuals who take at least one exam in the respective academic year. Controls include year dummies, an indicator for completion of the entrance exam (interacted with the treatment variable), an indicator for non-German mother tongue, and group size. Standard errors are computed using the delta-method and clustered at the group level.

Table 2.4: Average partial effects of a high share of female peers (coded as dummy variable) on academic performance – male students

(1) Dependent variables: Retention during the first year

	First semester		Second semester	
	completed	passed	completed	passed
Fraction female > 0.3	-0.01	-0.01	-0.01	-0.01
S.E.	(0.01)	(0.02)	(0.02)	(0.02)
p-value	(0.60)	(0.58)	(0.70)	(0.56)
Obs.	3,431	3,431	3,431	3,431
R-squared	0.05	0.02	0.02	0.02
Controls	Yes	Yes	Yes	Yes

(2) Dependent variables: Major choice

	Enters 2nd year	Business	Economics	International Affairs
Fraction female > 0.3	0.00	0.03	0.00	0.00
S.E.	(0.02)	(0.02)	(0.01)	(0.01)
p-value	(0.95)	(0.11)	(0.77)	(0.87)
Obs.	3,431	3,431	3,431	3,431
R-squared	0.03	0.01	0.04	0.01
Controls	Yes	Yes	Yes	Yes

(3) Dependent variables: Course grades, first semester

	Math	Economics	Business	Legal studies
Fraction female > 0.3	0.03	-0.01	-0.01	-0.01
S.E.	(0.04)	(0.04)	(0.04)	(0.04)
p-value	(0.47)	(0.73)	(0.86)	(0.74)
Obs.	3,244	3,375	3,398	3,335
R-squared	0.05	0.04	0.03	0.03
Controls	Yes	Yes	Yes	Yes

Continued on the next page.

Continued from the previous page: Average partial effects of a high share of female peers (coded as dummy variable) on academic performance – male students

(4) Dependent variables: Average grades, first four semesters				
	1st semester	2nd semester	3rd semester	4th semester
Fraction female > 0.3	0.00	-0.01	0.00	-0.02
S.E.	(0.04)	(0.04)	(0.04)	(0.05)
p-value	(0.95)	(0.82)	(0.98)	(0.63)
Obs.	3,418	3,418	2,675	2,721
R-squared	0.05	0.05	0.07	0.03
Controls	Yes	Yes	Yes	Yes

The table presents average partial effects of an above-average share of female students in the group (binary variable: share of female $>$ 30%) on performance indicators for male students (3,431 observations). The analysis uses probit models for binary performance outcomes (panels 1 and 2) and OLS models for continuous performance outcomes (panels 3 and 4). Grades are standardized at the course-cohort level (panel 3) or cohort level (panel 4). Grades are missing for dropouts and individuals who did not take the respective exam. Average grades are defined for all individuals who take at least one exam in the respective academic year. Controls include year dummies, an indicator for completion of the entrance exam (interacted with the treatment variable), an indicator for non-German mother tongue, and group size. Standard errors are computed using the delta-method and clustered at the group level. Source: Own calculations based on academic records from the University of St. Gallen.

facilitates the comparison of relative magnitudes. For female students, the dummy variable analysis suggests larger effects than the analysis based on a continuous variable, especially for those outcomes that are weakly significant in the dummy variable model. For male students, results of the two modeling approaches differ, but no clear tendency in the magnitude can be established. The lack of robustness with respect to this component of model choice shows that the peer effect estimates have to be interpreted with caution when it comes to policy recommendations.

2.6.3 Segregation effects

Table 2.5: Segregation effects: Course grades, first semester

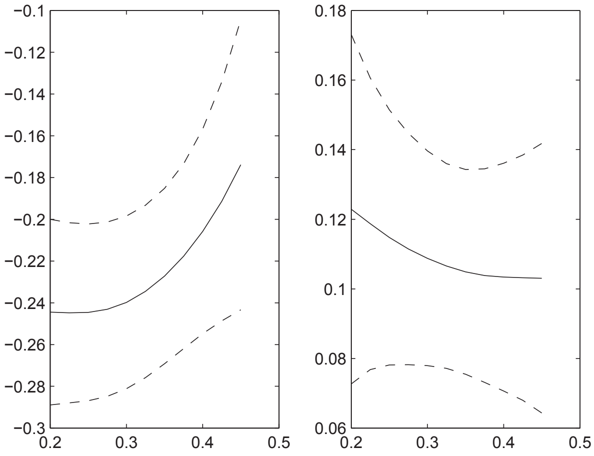
	ASE	LSOE	LPPE	LEPE	LSIE
Math	0.1230	0.0104	0.0009	0.0227	0.0329
S.E.	(0.1283)	(0.0111)	(0.0024)	(0.0117)	(0.0239)
p-values	(0.3377)	(0.3491)	(0.7134)	(0.0530)	(0.1681)
Sample	4,245 observations				
Economics	0.0624	0.0093	0.0007	0.0197	0.0395
S.E.	(0.1255)	(0.0106)	(0.0028)	(0.0115)	(0.0215)
p-values	(0.6190)	(0.3797)	(0.8017)	(0.0878)	(0.0663)
Sample	4,512 observations				
Business	0.0675	0.0065	0.0006	0.0110	0.0418
S.E.	(0.1244)	(0.0105)	(0.0028)	(0.0115)	(0.0230)
p-values	(0.5873)	(0.5364)	(0.8263)	(0.3413)	(0.0695)
Sample	4,546 observations				
Legal studies	0.0663	0.0067	0.0014	0.0116	0.0407
S.E.	(0.1208)	(0.0100)	(0.0027)	(0.0107)	(0.0232)
p-values	(0.5830)	(0.5007)	(0.6019)	(0.2802)	(0.0790)
Sample	4,451 observations				

The table shows gender segregation effects (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors and p-values are in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

This section on segregation effects focuses on first semester grades (Table 2.5); results for retention, major choice, and average grades are overall less robust (see Section 2.6.4). In line with the weak spillover effects discussed in the previous section, the average spillover effects are insignificant. Yet, Table 2.5 displays significant *local external peer effects* (LPPE) and *local segregation inequality effects* (LSIE). Segregation increases inequality in grades in economics, business, and math, and increases outcomes for the stayers both in math and in economics (significant at the 10%-level). The results are robust to the choice of the bandwidths (Table 2.A.15).

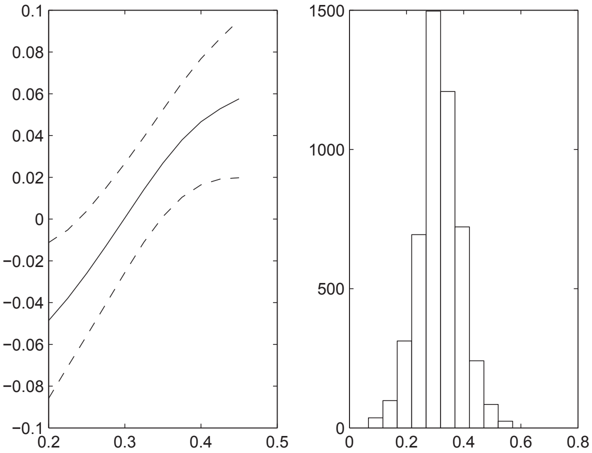
Figures and 2.2 and 2.3 display the results for economics grades; furthermore, I would like to illustrate the mechanisms at play with the following example. Suppose that only two groups of 10 individuals exist. One group has four female members (high-female group), and the other has three female members (low-female group). Segregation can be increased by switching a female member of the low-female group with a male member of the high-female group. As Figure 2.2 displays, females in the high-female group would benefit more from this increase in segregation than females in the low-female group would suffer (i.e., the slope at a share of 0.4 is steeper than the slope at a share of 0.3). Similarly, males in the low-female group would benefit more from this increase in segregation than males in the high-share group would suffer. Thus, the overall effect on the "stayers" is positive. The effect on the "movers" is also positive, as the female mover would benefit from moving from the low- to the high-female group, and the male mover would benefit, though only slightly, from moving from the high- to the low-female group. The average effect on the movers for economics grades, however, is not significant at any conventional level. The overall gain in outcomes from segregation, including both the movers and the stayers, is higher for females than for males, indicating that segregation will close the gender gap in economics grades to some extent. The positive local segregation outcome effect confirms this result (2.5): A positive coefficient implies a reduction in the negative gap between males and females.

Figure 2.2: Mean allocation response functions by gender – economics grade



The left panel shows the mean allocation response function for female students, the right panel shows the mean allocation response function for male students. The x-axis displays the share of female peers; the y-axis displays average economics grades. The estimation uses trimmming of observations with a share of female students below 0.2 and above 0.45. Estimation and inference closely follows Graham et al. (2010). For details, please also refer to Sections 2.5 and 2.B. Dashed lines: 90%-confidence bands. Based on 4,512 observations.

Figure 2.3: Mean allocation response function and distribution of the share of females



The left panel shows the mean allocation response function, averaged over males and females. The x-axis displays the share of female peers; the y-axis displays average economics grades. The right panel shows the distribution of the share of female students in a group (the estimation uses trimming of observations below 0.2 and above 0.45). Estimation and inference closely follows Graham et al. (2010). For details, please also refer to Sections 2.5 and 2.B. Dashed lines: 90%-confidence bands. Based on 4,512 observations.

2.6.4 Segregation effects: Robustness

I provide the following two robustness checks for the nonparametric analysis. First, I compare the average spillover effects from the nonparametric model to the average spillover effects implied by the parametric models discussed in Section 2.6.2. Second, I assess bandwidth choice by testing alternative bandwidths as suggested by Graham et al. (2010).

With respect to the first check, average spillover effects are only partly robust across parametric and nonparametric specifications. Results for first semester grades are particularly robust across specifications, both in terms of their quality and in terms of their magnitude (see Table 2.A.12).

Tables 2.A.13-2.A.16 present robustness checks with respect to bandwidth choices. The results for first semester grades are particularly robust, whereas the results for the other outcomes are not. The lack in robustness might come from two sources. First, binary outcomes such as retention and major choice are difficult to analyze in a nonparametric framework, due to low variation in these outcomes. Second, average grades might not generate meaningful variation. Averages are taken over all courses that a student completes, but some students only complete very few courses, especially in the first two semesters when many students drop out. These averages might not proxy performance well. Therefore, I do not draw further conclusions on segregation effects with respect to these outcomes.

2.7 Discussion

This paper analyses peer effects in higher education. The paper tests whether team composition during the freshmen week at the University of St. Gallen matters for subsequent academic outcomes such as retention, major choice, or grades. I assess gender group composition under the hypothesis that higher shares of female peers might help female students to find friends and study partners, owing to homophily in social network formation. To assess the existence of peer and reallocation effects, I apply a set of parametric models as well as the nonparametric framework by Graham et al. (2010) in order to investigate the existence of beneficial reallocations.

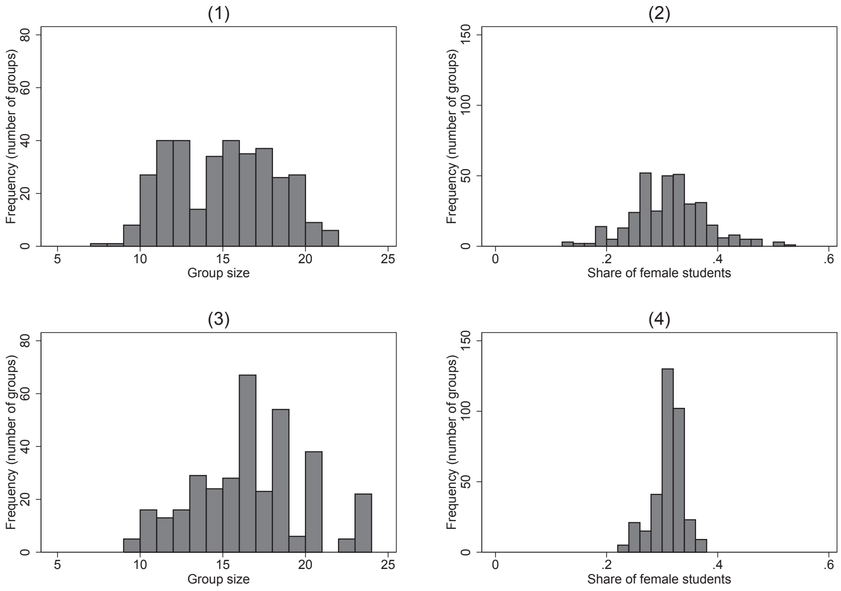
In line with the existing literature (Oosterbeek and van Ewijk, 2014), gender peer effects in higher education are rather small. Yet, in contrast to Oosterbeek and van Ewijk (2014), who find small results for males and no results for females, this paper finds small results for females and no results for males. In particular, females' math grades and females' probability of passing the first year seem positively affected (significant at the 10%-level). Overall, the estimation results suffer from a lack in precision.

Furthermore, small effects of segregation on first-semester grades exist (significant at the 10%-level). In particular, moving to a more segregated allocation would improve grades for some individuals, and in particular for females, thus locally reducing inequality in outcomes. Given the lack of precision in these results, however, this paper abstains from any policy recommendations. Further research on segregation effects is needed to complement this study.

Appendix

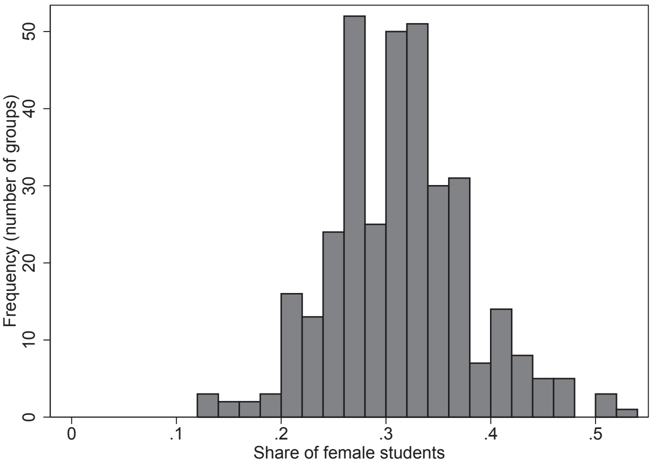
2.A Figures and tables

Figure 2.A.1: Variation of group sizes and the fraction of male students between freshmen groups as a result of the randomization procedure



The figure shows a comparison of group size (panels 1 and 3 to the left) and the share of female students (panels 2 and 3 to the right) between the original assignment to freshmen groups (panels 1 and 2 above) and a simulated assignment according to the assignment rules stated by the university (panels 3 and 4 below). The number of groups in each year in the simulation is set to the number of groups in each year in the original data. The simulation applies the stratificiation scheme of the original assignment procedure: The university administration defines the quasi-random group allocation on the basis of last names, conditional on gender and whether students are admitted based on an entrance exam, which is mandatory for foreign students (“admission rule”). First, students are split into four strata (s), according to gender and admission rule. Second, within strata, students are ranked according to the position of their last name in the alphabet. Consider k teams. Within each stratum s, student of rank 1 is assigned to team 1, student of rank 2 is assigned to team 2, and so forth, until all k teams have one member; student of rank $k + 1$ is assigned to team 1, student of rank $k + 2$ is assigned to team 2 and so forth. Subsequently, the university administration assigns each team to a room. If rooms are too small, the administration randomly redistributes the teams so that group and room size match. The administration assigns tutors on the basis of the second name as well, according to the procedure mentioned above. The quasi-random procedure leads to groups of different sizes and creates variation in the gender share.

Figure 2.A.2: Distribution of the fraction of females in each group



The histogram shows the distribution of the fraction of females across groups for the pooled sample (345 groups). Maximum value: 0.07; minimum value: 0.57; mean: 0.32; standard deviation: 0.07. Source: Own calculation based on administrative data from the University of St. Gallen.

Table 2.A.1: Sample construction, resulting number of groups and group size

	Total	2003	2004	2006	2007	2008	2009
1st year students	5,204	699	636	812	905	1,102	1,050
Estimation sample	5,012	596	631	806	865	1,082	1,032
(% of 1st year students)	96%	85%	99%	99%	96%	98%	98%
# of groups	345	56	56	57	56	60	60
Group size							
Mean	15	11	11	14	15	18	17
Median	15	11	12	14	15	18	17
Smallest	7	8	7	10	12	15	11
Biggest	22	12	13	16	18	22	21

The table shows the fraction of students included in the estimation sample, out of all students who enter the first year, for cohorts 2003-2009. Cohort 2005 is excluded due to missing information on freshmen groups. The following individuals are excluded form the estimation sample: Individuals who did not participate in the freshmen week, who signed up for special groups (e.g. media group), or who were in groups with a fraction of males of 100%. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.2: Descriptive statistics: Student characteristics, group variables, and outcomes by year

Year	Total	2003	2004	2006	2007	2008	2009
Background characteristics							
Age	20	20	20	20	20	20	20
Female	0.32	0.32	0.29	0.31	0.31	0.33	0.33
Entrance exam	0.18	0.22	0.17	0.15	0.18	0.19	0.20
Non-Swiss nationality	0.24	0.29	0.23	0.20	0.25	0.25	0.25
Non-German mothertongue	0.11	0.10	0.13	0.10	0.09	0.11	0.12
High school St. Gallen	0.15	0.18	0.18	0.15	0.15	0.15	0.13
Extended track	0.06	0.02	0.05	0.05	0.05	0.07	0.08
Group variables							
Share female	0.32	0.32	0.29	0.31	0.31	0.33	0.33
Group size	15.20	10.72	11.38	14.26	15.57	18.18	17.43
Outcomes							
1st semester completed	0.93	0.83	0.96	0.94	0.94	0.94	0.93
1st semester passed	0.82	0.82	0.82	0.82	0.82	0.83	0.80
2nd semester completed	0.75	0.79	0.77	0.79	0.74	0.73	0.72
2nd semester passed	0.66	0.69	0.68	0.71	0.64	0.63	0.64
Enters 2nd year	0.79	0.81	0.79	0.81	0.78	0.78	0.76
Business	0.52	0.53	0.51	0.54	0.54	0.51	0.51
Economics	0.12	0.09	0.12	0.13	0.11	0.13	0.13
International Affairs	0.11	0.13	0.11	0.11	0.10	0.11	0.09
Law & Econ.	0.05	0.04	0.06	0.06	0.05	0.04	0.05
Legal studies	0.05	0.07	0.08	0.05	0.04	0.04	0.03
Grade math	0.00	0.01	0.00	0.01	0.00	0.01	0.00
Grade economics	0.00	-0.01	0.01	0.00	0.01	0.01	0.00
Grade business	0.01	0.01	0.01	0.00	0.01	0.01	0.00
Grade legal studies	0.00	-0.01	0.01	0.00	0.01	0.01	0.00
GPA semester 1	0.01	0.01	0.01	0.00	0.01	0.01	0.00
GPA semester 2	0.00	0.01	0.01	0.00	0.01	0.01	0.00
GPA semester 3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPA semester 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The table shows mean student characteristic, group variables, and outcome variables by cohort (year of entry) for the estimation sample (5,012 observations). "Share female" is the treatment variable and computed as the leave-own-out mean. Course grades are standardized at the course level. Grades are missing for dropouts and for individuals who did not take the respective exam. Average grades (GPA) are defined for all individuals who take at least one exam in the respective academic years and standardized at the cohort level. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.3: Gender gap in the probability of passing – estimation sample (probit model)

Dependent variable: probability of passing	APE	S.E.	APE	S.E.
Female	-0.08***	(0.01)	-0.06***	(0.01)
Entrance Exam			0.26***	(0.02)
Age			-0.02***	(0.00)
Non-Swiss nationality			-0.13***	(0.03)
Non-German mother tongue			-0.16***	(0.03)
High school St. Gallen			-0.01	(0.02)
Extended track			-0.02	(0.04)
Year dummies	No		Yes	
Pseudo R2	0.0046		0.0467	
Observations	5,012		5,012	

The table shows average partial effects, computed from a probit regression of individual students characteristics on the probability of passing the first year in the first attempt (5,012 observations). Standard errors based on the delta method are in parentheses.*Significant at the 10%-level, **significant at the 5%-level, ***significant at the 1%-level.

Table 2.A.4: Gender gap in the probability of passing – cohort 2011 (probit model)

Dependent variable: probability of passing	APE	S.E.	APE	S.E.
Female	-0.09***	(0.03)	-0.10***	(0.00)
Entrance test	0.16***	(0.02)	0.16***	(0.00)
Older than 20 years	-0.06**	(0.03)	-0.02	(0.40)
Non-German mother tongue	-0.02	(0.06)	-0.05	(0.48)
Legal studies track	-0.04	(0.05)	-0.02	(0.59)
High school grade			0.10***	(0.00)
Pseudo R2	0.0513		0.119	

Sample: 958 observations. Cohort 2011. Comparison of average marginal effects from a logit regression of individual characteristics on the probability of passing, with and without high school GPA (standardized at the high school country level) as a measure of ability. Standard errors based on the delta method in parenthesis. *Significant at the 10%-level, **significant at the 5%-level, ***significant at the 1%-level.

Table 2.A.5: Average marginal effect of the share of female peers (third order polynomial) on academic performance – female students

(1) Dependent variables: Retention during the first year

	First semester		Second semester	
	completed	passed	completed	passed
Share female	0.12	0.07	0.10	0.10
S.E.	(0.10)	(0.16)	(0.17)	(0.18)
p-value	(0.21)	(0.67)	(0.53)	(0.56)
Obs.	1,581	1,581	1,581	1,581
R-squared	0.04	0.01	0.02	0.04
Controls	Yes	Yes	Yes	Yes

(2) Dependent variables: Major choice

	Enters 2nd year	Business	Economics	Interational Affairs
Share female	0.08	0.03	0.07	0.21
S.E.	(0.15)	(0.18)	(0.10)	(0.13)
p-value	(0.59)	(0.89)	(0.48)	(0.10)
Obs.	1,581	1,581	1,581	1,581
R-squared	0.02	0.02	0.04	0.01
Controls	Yes	Yes	Yes	Yes

(3) Dependent variables: Course grades, first semester

	Math	Economics	Business	Legal studies
Share female	0.52	0.38	0.23	0.50
S.E.	(0.35)	(0.32)	(0.33)	(0.34)
p-value	(0.14)	(0.23)	(0.49)	(0.14)
Obs.	1,375	1,541	1,554	1,517
R-squared	0.05	0.05	0.04	0.04
Controls	Yes	Yes	Yes	Yes

Continued on the next page.

Continued from the previous page: Average marginal effect of the share of female peers (third order polynomial) on academic performance – female students

(4) Dependent variables: Average grades, first four semesters				
	1st semester	2nd semester	3rd semester	4th semester
Share female	0.40	0.32	-0.19	-0.19
S.E.	(0.32)	(0.33)	(0.38)	(0.39)
p-value	(0.21)	(0.33)	(0.62)	(0.63)
Obs.	1,563	1,563	1,102	1,120
R-squared	0.06	0.05	0.06	0.04
Controls	Yes	Yes	Yes	Yes

The table presents average marginal effects of the share of female students in the group (third order polynomial) on performance indicators for female students (1,581 observations), using a third-order polynomial for the share of females as independent variable. The analysis uses probit models for regressions on binary performance indicators (panels 1 and 2) and OLS models for regressions on continuous performance indicators (panels 3 and 4). Grades are standardized at the course-cohort level (panel 3) or cohort level (panel 4). Grades are missing for dropouts and individuals who did not take the respective exam. Average grades are defined for all individuals who take at least one exam in the respective academic year. Controls include year dummies, an indicator for completion of the entrance exam (interacted with the treatment variable), an indicator for non-German mother tongue, and group size. Standard errors are computed using the delta-method and clustered at the group level. Source: Own calculations based on academic records from the University of St. Gallen.

Table 2.A.6: Average marginal effect of the share of female peers (third order polynomial) on academic performance – male students

(1) Dependent variables: Retention during the first year

	First semester		Second semester	
	completed	passed	completed	passed
Share female	-0.04	-0.12	-0.12	-0.13
S.E.	(0.06)	(0.10)	(0.11)	(0.12)
p-value	(0.51)	(0.22)	(0.29)	(0.26)
Obs.	3,431	3,431	3,431	3,431
R-squared	0.05	0.02	0.02	0.03
Controls	Yes	Yes	Yes	Yes

(2) Dependent variables: Major choice

	Enters 2nd year	Business	Economics	Interational Affairs
Share female	-0.06	0.17	-0.02	-0.06
S.E.	(0.10)	(0.13)	(0.08)	(0.07)
p-value	(0.52)	(0.20)	(0.82)	(0.39)
Obs.	3,431	3,431	3,431	3,431
R-squared	0.03	0.01	0.04	0.01
Controls	Yes	Yes	Yes	Yes

(3) Dependent variables: Course grades, first semester

	Math	Economics	Business	Legal studies
Share female	0.03	-0.10	0.01	-0.16
S.E.	(0.28)	(0.24)	(0.26)	(0.24)
p-value	(0.91)	(0.68)	(0.98)	(0.53)
Obs.	3,244	3,375	3,398	3,335
R-squared	0.05	0.04	0.03	0.03
Controls	Yes	Yes	Yes	Yes

Continued on the next page.

Continued from the previous page: Average marginal effect of the share of female peers (third order polynomial) on academic performance – male students

(4) Dependent variables: Average grades, first four semesters				
	1st semester	2nd semester	3rd semester	4th semester
Share female	-0.09	-0.10	-0.01	-0.10
S.E.	(0.25)	(0.26)	(0.27)	(0.30)
p-value	(0.71)	(0.69)	(0.98)	(0.75)
Obs.	3,418	3,418	2,675	2,721
R-squared	0.05	0.05	0.07	0.03
Controls	Yes	Yes	Yes	Yes

The table presents average marginal effects of the share of female students in the group (third order polynomial) on performance indicators for male students (2,445 observations), using a third-order polynomial for the share of females as independent variable. The analysis uses probit models for regressions on binary performance indicators (panels 1 and 2) and OLS models for regressions on continuous performance indicators (panels 3 and 4). Grades are standardized at the course-cohort level (panel 3) or cohort level (panel 4). Grades are missing for dropouts and individuals who did not take the respective exam. Average grades are defined for all individuals who take at least one exam in the respective academic year. Controls include year dummies, an indicator for completion of the entrance exam (interacted with the treatment variable), an indicator for non-German mother tongue, and group size. Standard errors are computed using the delta-method and clustered at the group level. Source: Own calculations based on academic records from the University of St. Gallen.

Table 2.A.7: Segregation effects: Retention during the first year

	ASE	LSOE	LPPE	LEPE	LSIE
1st semester completed	-0.0007	0.0005	0.0002	0.0005	-0.0067
S.E.	(0.0354)	(0.0031)	(0.0043)	(0.0053)	(0.0388)
p-values	(0.9832)	(0.8775)	(0.9590)	(0.9243)	(0.8626)
Sample			4,707 observations		
1st semester passed	0.0408	-0.0086	0.0002	-0.0312	0.0503
S.E.	(0.0508)	(0.0047)	(0.0039)	(0.0070)	(0.0344)
p-values	(0.4226)	(0.0655)	(0.9600)	(0.0000)	(0.1440)
Sample			4,707 observations		
2nd semester completed	0.0019	-0.0007	-0.0002	-0.0012	-0.0064
S.E.	(0.0728)	(0.0057)	(0.0019)	(0.0064)	(0.0216)
p-values	(0.9793)	(0.9085)	(0.9175)	(0.8476)	(0.7687)
Sample			4,707 observations		
2nd semester passed	0.0165	0.0049	0.0009	0.0078	0.0158
S.E.	(0.0621)	(0.0051)	(0.0033)	(0.0061)	(0.0291)
p-values	(0.7907)	(0.3367)	(0.7818)	(0.2016)	(0.5858)
Sample			4,707 observations		

The table shows gender segregation effects (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors and p-values in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.8: Segregation effects: Major choice

	ASE	LSOE	LPPE	LEPE	LSIE
Enters 2nd year	0.0180	0.0018	0.0003	0.0029	0.0030
S.E.	(0.0664)	(0.0052)	(0.0037)	(0.0063)	(0.0462)
p-values	(0.7868)	(0.7272)	(0.9290)	(0.6464)	(0.9489)
Sample			4,707 observations		
Major: Business	0.1153	0.0005	-0.0004	0.0114	0.0060
S.E.	(0.0661)	(0.0057)	(0.0026)	(0.0062)	(0.0228)
p-values	(0.0813)	(0.9314)	(0.8863)	(0.0635)	(0.7928)
Sample			4,707 observations		
Major: Econ	0.0245	-0.0016	-0.0000	0.0010	0.0041
S.E.	(0.0400)	(0.0034)	(0.0008)	(0.0036)	(0.0074)
p-values	(0.5402)	(0.6342)	(0.9845)	(0.7805)	(0.5809)
Sample			4,707 observations		
Major: Int. Affairs	0.0021	0.0006	0.0003	0.0003	0.0018
S.E.	(0.0399)	(0.0033)	(0.0010)	(0.0036)	(0.0097)
p-values	(0.9574)	(0.8541)	(0.7253)	(0.9257)	(0.8530)
Sample			4,707 observations		

The table shows gender segregation effects (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors and p-values in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.9: Segregation effects: Average grades, first four semesters

	ASE	LSOE	LPPE	LEPE	LSIE
1st semester	0.2819	-0.0184	0.0007	-0.0739	0.1793
S.E.	(0.1188)	(0.0108)	(0.0028)	(0.0120)	(0.0236)
p-values	(0.0177)	(0.0899)	(0.7977)	(0.0000)	(0.0000)
Sample	4,572 observations				
2nd semester	0.2978	-0.0247	0.0005	-0.0207	0.0931
S.E.	(0.1208)	(0.0111)	(0.0028)	(0.0119)	(0.0232)
p-values	(0.0137)	(0.0254)	(0.8698)	(0.0833)	(0.0001)
Sample	4,572 observations				
3rd semester	0.0184	-0.0007	-0.0002	0.0045	0.0106
S.E.	(0.1345)	(0.0121)	(0.0026)	(0.0126)	(0.0240)
p-values	(0.8912)	(0.9562)	(0.9499)	(0.7206)	(0.6594)
Sample	3,466 observations				
4th semester	-0.0090	-0.0050	-0.0004	-0.0069	0.0235
S.E.	(0.1389)	(0.0120)	(0.0023)	(0.0123)	(0.0257)
p-values	(0.9481)	(0.6766)	(0.8532)	(0.5724)	(0.3589)
Sample	3,523 observations				

The table shows gender segregation effects (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors and p-values in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.10: Robustness: Comparison across models – female students

(1) Dependent variables: Retention during the first year				
Model	First semester		Second semester	
	completed	passed	completed	passed
Dummy variable	0.17	0.12	0.31	0.44
First order	0.13	0.10	0.13	0.11
Second order	0.13	0.10	0.12	0.09
Third order	0.12	0.07	0.10	0.10
(2) Dependent variables: Major choice				
Model	Enters 2nd year	Business	Economics	Interational Affairs
Dummy variable	0.26	0.01	0.08	0.34
First order	0.06	0.03	0.07	0.20
Second order	0.05	0.02	0.07	0.20
Third order	0.08	0.03	0.07	0.21
(3) Dependent variables: Course grades, first semester				
Model	Math	Economics	Business	Legal studies
Dummy variable	0.84	0.17	0.41	0.40
First order	0.50	0.51	0.30	0.54
Second order	0.50	0.50	0.29	0.55
Third order	0.52	0.38	0.23	0.50
(4) Dependent variables: Average grades, first four semesters				
Model	1st semester	2nd semester	3rd semester	4th semester
Dummy variable	0.54	0.53	-0.70	-0.17
First order	0.46	0.37	-0.12	-0.11
Second order	0.45	0.36	-0.11	-0.10
Third order	0.40	0.32	-0.19	-0.19

The table compares average partial effects of marginal increases in the share of female students across different models, for female students (1,581 observations). The dummy variable takes the value of 1 if the share of female students is above 0.3. To convert the estimate into a partial effect, I divide the coefficient by 0.11, which is the average difference in the share of females between groups with a value 0 and a value of 1 for the variable. The other estimates rely on parametric specifications as outlined in Section 2.5.

Table 2.A.11: Robustness: Comparison across models – male students

(1) Dependent variables: Retention during the first year				
Model	First semester		Second semester	
	completed	passed	completed	passed
Dummy variable	-0.05	-0.08	-0.06	-0.10
First order	-0.04	-0.11	-0.11	-0.10
Second order	-0.04	-0.11	-0.11	-0.10
Third order	-0.04	-0.12	-0.12	-0.13
(2) Dependent variables: Major choice				
Model	Enters 2nd year	Business	Economics	Interational Affairs
Dummy variable	-0.05	-0.08	-0.06	-0.10
First order	-0.04	-0.11	-0.11	-0.10
Second order	-0.04	-0.11	-0.11	-0.10
Third order	-0.04	-0.12	-0.12	-0.13
(3) Dependent variables: Course grades, first semester				
Model	Math	Economics	Business	Legal studies
Dummy variable	0.28	-0.12	-0.07	-0.12
First order	0.00	-0.11	0.02	-0.16
Second order	0.00	-0.11	0.02	-0.16
Third order	0.03	-0.10	0.01	-0.16
(4) Dependent variables: Average grades, first four semesters				
Model	1st semester	2nd semester	3rd semester	4th semester
Dummy variable	-0.02	-0.08	0.01	-0.21
First order	-0.10	-0.09	-0.02	-0.12
Second order	-0.10	-0.09	-0.02	-0.12
Third order	-0.09	-0.10	-0.01	-0.10

The table compares average partial effects of marginal increases in the share of female students across different models, for female students (3,431 observations). The dummy variable takes the value of 1 if the share of female students is above 0.3. To convert the estimate into a partial effect, I divide the coefficient by 0.11, which is the average difference in the share of females between groups with a value 0 and a value of 1 for the variable. The other estimates rely on parametric specifications as outlined in Section 2.5.

Table 2.A.12: Robustness: Comparison across models – all students

(1) Dependent variables: Retention during the first year

Model	First semester		Second semester	
	completed	passed	completed	passed
Dummy variable	0.00	0.00	0.01	0.01
First order	0.02	-0.04	-0.03	-0.03
Second order	0.01	-0.05	-0.03	-0.05
Third order	0.01	-0.06	-0.04	-0.05
Nonparametric	0.00	0.04	0.00	0.02

(2) Dependent variables: Major choice

Model	Enters 2nd year	Business	Economics	International Affairs
Dummy variable	0.00	0.00	0.01	0.01
First order	0.02	-0.04	-0.03	-0.03
Second order	0.01	-0.05	-0.03	-0.05
Third order	0.01	-0.06	-0.04	-0.05
Nonparametric	0.00	0.04	0.00	0.02

(3) Dependent variables: Course grades, first semester

Model	Math	Economics	Business	Legal studies
Dummy variable	0.05	0.00	0.01	0.01
First order	0.16	0.10	0.11	0.07
Second order	0.14	0.08	0.11	0.08
Third order	0.19	0.06	0.08	0.06
Nonparametric	0.12	0.06	0.07	0.07

(4) Dependent variables: Average grades, first four semesters

Model	1st semester	2nd semester	3rd semester	4th semester
Dummy variable	0.02	0.01	-0.02	-0.02
First order	0.09	0.07	-0.05	-0.12
Second order	0.09	0.06	-0.05	-0.09
Third order	0.07	0.04	-0.06	-0.13
Nonparametric	0.28	0.30	0.02	0.00

The table compares average partial effects of marginal increases in the share of female students ("average spillover effects") across different models, for the full estimation sample (5,012 observations for parametric and 4,707 observations for nonparametric specifications). The dummy variable takes the value of 1 if the share of female students is above 0.3. To convert the estimate into a partial effect, I divide the coefficient by 0.11, which is the average difference in the share of females between groups with a value 0 and a value of 1 for the variable. For the estimates of parametric and nonparametric specifications, see Section 2.5.

Table 2.A.13: Robustness: Retention during the first year

	ASE	LSOE	LPPE	LEPE	LSIE
			First semester completed		
CV	-0.0007	0.0005	0.0002	0.0005	-0.0067
S.E.	(0.0354)	(0.0031)	(0.0043)	(0.0053)	(0.0388)
5/6 of CV	0.0007	0.0007	0.0003	0.0009	-0.0057
S.E.	(0.0365)	(0.0030)	(0.0044)	(0.0053)	(0.0408)
2/3 of CV	0.0018	0.0008	0.0004	0.0011	-0.0051
S.E.	(0.0385)	(0.0031)	(0.0044)	(0.0054)	(0.0460)
1/2 of CV	0.0034	0.0002	0.0005	0.0011	-0.0036
S.E.	(0.0426)	(0.0035)	(0.0044)	(0.0056)	(0.0587)
			First semester passed		
CV	0.0408	-0.0086	0.0002	-0.0312	0.0503
S.E.	(0.0508)	(0.0047)	(0.0039)	(0.0070)	(0.0344)
5/6 of CV	0.0751	-0.0116	0.0002	-0.1186	0.1818
S.E.	(0.0545)	(0.0050)	(0.0039)	(0.0109)	(0.0348)
2/3 of CV	0.1165	-0.0154	0.0002	-2.6922	3.9581
S.E.	(0.0612)	(0.0057)	(0.0039)	(0.1460)	(0.2047)
1/2 of CV	0.1364	-0.0150	0.0002	-4013.1900	5880.2340
S.E.	(0.0739)	(0.0070)	(0.0039)	(212.8011)	(320.1824)
			Second semester completed		
CV	0.0019	-0.0007	-0.0002	-0.0012	-0.0064
S.E.	(0.0728)	(0.0057)	(0.0019)	(0.0064)	(0.0216)
5/6 of CV	0.0008	-0.0011	-0.0003	-0.0023	-0.0104
S.E.	(0.0757)	(0.0057)	(0.0019)	(0.0064)	(0.0226)
2/3 of CV	-0.0030	-0.0019	-0.0004	-0.0035	-0.0168
S.E.	(0.0810)	(0.0060)	(0.0019)	(0.0065)	(0.0256)
1/2 of CV	-0.0107	-0.0027	-0.0006	-0.0042	-0.0253
S.E.	(0.0895)	(0.0071)	(0.0019)	(0.0073)	(0.0316)

Continued on the next page.

Continued from the previous page: Robustness: Retention during the first year

	ASE	LSOE	LPPE	LEPE	LSIE
	Second semester passed				
CV	0.0165	0.0049	0.0009	0.0078	0.0158
S.E.	(0.0621)	(0.0051)	(0.0033)	(0.0061)	(0.0291)
Estimate	0.0209	0.0051	0.0010	0.0080	0.0191
2/3 of CV	(0.0651)	(0.0052)	(0.0033)	(0.0062)	(0.0304)
Estimate	0.0310	0.0047	0.0012	0.0084	0.0246
S.E.	(0.0699)	(0.0054)	(0.0033)	(0.0065)	(0.0340)
1/2 of CV	0.0424	0.0038	0.0013	0.0090	0.0279
S.E.	(0.0784)	(0.0061)	(0.0033)	(0.0071)	(0.0430)

The table shows robustness of gender segregation effects with respect to different bandwidths, defined relative to the cross-validation (CV) bandwidth (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors are in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.14: Robustness: Major choice

	ASE	LSOE	LPPE	LEPE	LSIE
			2nd year started		
CV	0.0180	0.0018	0.0003	0.0029	0.0030
S.E.	(0.0664)	(0.0052)	(0.0037)	(0.0063)	(0.0462)
5/6 of CV	0.0277	0.0027	0.0004	0.0035	0.0053
S.E.	(0.0705)	(0.0053)	(0.0037)	(0.0064)	(0.0495)
2/3 of CV	0.0413	0.0039	0.0006	0.0039	0.0068
S.E.	(0.0783)	(0.0060)	(0.0037)	(0.0070)	(0.0560)
1/2 of CV	0.0593	0.0051	0.0008	0.0041	0.0083
S.E.	(0.0875)	(0.0072)	(0.0037)	(0.0081)	(0.0664)
			Major: Business		
CV	0.1153	0.0005	-0.0004	0.0114	0.0060
S.E.	(0.0661)	(0.0057)	(0.0026)	(0.0062)	(0.0228)
5/6 of CV	0.1287	-0.0005	-0.0003	0.0064	0.0134
S.E.	(0.0686)	(0.0057)	(0.0026)	(0.0063)	(0.0236)
2/3 of CV	0.1561	-0.0026	-0.0003	-0.0373	0.0747
S.E.	(0.0725)	(0.0061)	(0.0026)	(0.0075)	(0.0256)
1/2 of CV	0.2044	-0.0060	-0.0003	-1.9835	2.9215
S.E.	(0.0801)	(0.0071)	(0.0026)	(0.1073)	(0.1530)
			Major: Economics		
CV	0.0245	-0.0016	-0.0000	0.0010	0.0041
S.E.	(0.0400)	(0.0034)	(0.0008)	(0.0036)	(0.0074)
5/6 of CV	0.0273	-0.0015	0.0000	0.0058	-0.0046
S.E.	(0.0426)	(0.0035)	(0.0008)	(0.0036)	(0.0078)
2/3 of CV	0.0294	-0.0015	0.0001	0.0297	-0.0427
S.E.	(0.0463)	(0.0036)	(0.0008)	(0.0041)	(0.0093)
1/2 of CV	0.0339	-0.0007	0.0001	0.7191	-1.0521
S.E.	(0.0522)	(0.0042)	(0.0008)	(0.0381)	(0.0602)

Continued on the next page.

Continued from the previous page: Robustness: Major choice

	ASE	LSOE	LPPE	LEPE	LSIE
		Major: International Affairs			
CV	0.0021	0.0006	0.0003	0.0003	0.0018
S.E.	(0.0399)	(0.0033)	(0.0010)	(0.0036)	(0.0097)
5/6 of CV	-0.0003	0.0007	0.0004	-0.0001	0.0023
S.E.	(0.0418)	(0.0034)	(0.0010)	(0.0037)	(0.0101)
2/3 of CV	-0.0037	0.0008	0.0005	-0.0012	0.0036
S.E.	(0.0447)	(0.0036)	(0.0010)	(0.0039)	(0.0109)
Estimate	-0.0058	0.0006	0.0006	-0.0035	0.0057
1/2 of CV	(0.0497)	(0.0041)	(0.0010)	(0.0044)	(0.0128)

The table shows robustness of gender segregation effects with respect to different bandwidths, defined relative to the cross-validation (CV) bandwidth (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors are in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.15: Robustness: Course grades, first semester

	ASE	LSOE	LPPE	LEPE	LSIE
			Math		
CV	0.1230	0.0104	0.0009	0.0227	0.0329
S.E.	(0.1283)	(0.0111)	(0.0024)	(0.0117)	(0.0239)
5/6 of CV	0.1308	0.0083	0.0011	0.0211	0.0295
S.E.	(0.1342)	(0.0112)	(0.0024)	(0.0119)	(0.0247)
2/3 of CV	0.1476	0.0068	0.0014	0.0216	0.0350
S.E.	(0.1443)	(0.0118)	(0.0024)	(0.0125)	(0.0274)
1/2 of CV	0.1730	0.0064	0.0016	0.0255	0.0464
S.E.	(0.1641)	(0.0136)	(0.0024)	(0.0143)	(0.0344)
			Economics		
CV	0.0624	0.0093	0.0007	0.0197	0.0395
S.E.	(0.1255)	(0.0106)	(0.0028)	(0.0115)	(0.0215)
5/6 of CV	0.0621	0.0087	0.0010	0.0188	0.0477
S.E.	(0.1317)	(0.0107)	(0.0028)	(0.0116)	(0.0214)
2/3 of CV	0.0694	0.0085	0.0013	0.0200	0.0620
S.E.	(0.1410)	(0.0110)	(0.0028)	(0.0119)	(0.0219)
1/2 of CV	0.0950	0.0092	0.0017	0.0248	0.0858
S.E.	(0.1580)	(0.0123)	(0.0028)	(0.0131)	(0.0254)
			Business		
CV	0.0675	0.0065	0.0006	0.0110	0.0418
S.E.	(0.1244)	(0.0105)	(0.0028)	(0.0115)	(0.0230)
5/6 of CV	0.0727	0.0060	0.0008	0.0103	0.0483
S.E.	(0.1301)	(0.0106)	(0.0028)	(0.0116)	(0.0230)
2/3 of CV	0.0907	0.0050	0.0011	0.0104	0.0579
S.E.	(0.1388)	(0.0110)	(0.0028)	(0.0119)	(0.0242)
1/2 of CV	0.1291	0.0046	0.0013	0.0132	0.0683
S.E.	(0.1528)	(0.0123)	(0.0028)	(0.0130)	(0.0287)

Continued on the next page.

Continued from the previous page: Robustness: Course grades, first semester

	ASE	LSOE	LPPE	LEPE	LSIE
Legal studies					
CV	0.0663	0.0067	0.0014	0.0116	0.0407
S.E.	(0.1208)	(0.0100)	(0.0027)	(0.0107)	(0.0232)
5/6 of CV	0.0688	0.0053	0.0016	0.0102	0.0423
S.E.	(0.1279)	(0.0101)	(0.0027)	(0.0108)	(0.0231)
2/3 of CV	0.0873	0.0034	0.0019	0.0093	0.0495
S.E.	(0.1395)	(0.0107)	(0.0027)	(0.0114)	(0.0240)
1/2 of CV	0.1166	0.0018	0.0022	0.0099	0.0546
S.E.	(0.1591)	(0.0125)	(0.0027)	(0.0131)	(0.0284)

The table shows robustness of gender segregation effects with respect to different bandwidths, defined relative to the cross-validation (CV) bandwidth (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors are in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

Table 2.A.16: Robustness: Average grades, first four semesters

	ASE	LSOE	LPPE	LEPE	LSIE
1st semester					
CV	0.2819	-0.0184	0.0007	-0.0739	0.1793
S.E.	(0.1188)	(0.0108)	(0.0028)	(0.0120)	(0.0236)
5/6 of CV	0.3731	-0.0246	0.0008	-0.5522	0.8804
S.E.	(0.1292)	(0.0117)	(0.0028)	(0.0308)	(0.0495)
2/3 of CV	0.4851	-0.0294	0.0009	-19.0752	27.9625
S.E.	(0.1472)	(0.0134)	(0.0028)	(1.0098)	(1.5227)
1/2 of CV	0.5227	-0.0209	0.0010	-28160.4833	41153.5670
S.E.	(0.1774)	(0.0164)	(0.0028)	(1492.0203)	(2248.0763)
2nd semester					
CV	0.2978	-0.0247	0.0005	-0.0207	0.0931
S.E.	(0.1208)	(0.0111)	(0.0028)	(0.0119)	(0.0232)
5/6 of CV	0.3970	-0.0327	0.0005	-0.2035	0.3597
S.E.	(0.1312)	(0.0121)	(0.0028)	(0.0159)	(0.0283)
2/3 of CV	0.5116	-0.0396	0.0005	-8.4686	12.4426
S.E.	(0.1502)	(0.0141)	(0.0028)	(0.4476)	(0.6748)
1/2 of CV	0.5278	-0.0319	0.0005	-12794.4022	18697.6874
S.E.	(0.1845)	(0.0180)	(0.0028)	(677.8824)	(1021.3971)
3rd semester					
CV	0.0184	-0.0007	-0.0002	0.0045	0.0106
S.E.	(0.1345)	(0.0121)	(0.0026)	(0.0126)	(0.0240)
5/6 of CV	0.0032	-0.0064	-0.0003	-0.0002	0.0063
S.E.	(0.1412)	(0.0124)	(0.0026)	(0.0129)	(0.0245)
2/3 of CV	-0.0178	-0.0135	-0.0006	-0.0059	0.0016
S.E.	(0.1520)	(0.0132)	(0.0026)	(0.0138)	(0.0257)
1/2 of CV	-0.0358	-0.0174	-0.0009	-0.0087	-0.0028
S.E.	(0.1696)	(0.0154)	(0.0026)	(0.0159)	(0.0292)

Continued on the next page.

Continued from the previous page: Robustness: Average grades, first four semesters

	ASE	LSOE	LPPE	LEPE	LSIE
			4th semester		
CV	-0.0090	-0.0050	-0.0004	-0.0069	0.0235
S.E.	(0.1389)	(0.0120)	(0.0023)	(0.0123)	(0.0257)
5/6 of CV	-0.0036	-0.0098	-0.0003	-0.0100	0.0316
S.E.	(0.1441)	(0.0120)	(0.0023)	(0.0124)	(0.0258)
2/3 of CV	-0.0000	-0.0178	-0.0004	-0.0151	0.0368
S.E.	(0.1522)	(0.0125)	(0.0023)	(0.0129)	(0.0267)
1/2 of CV	0.0144	-0.0252	-0.0007	-0.0184	0.0493
S.E.	(0.1649)	(0.0142)	(0.0023)	(0.0146)	(0.0293)

The table shows robustness of gender segregation effects with respect to different bandwidths, defined relative to the cross-validation (CV) bandwidth (ASE: average segregation effect, LSOE: local segregation outcome effect, LPPE: local private peer effect, LEPE: local external peer effect, LSIE: local segregation inequality effect). The sample is restricted to all individuals with a fraction of female freshmen peers between 20% and 45%. An indicator variable for completion of the entrance test is included in the regression as control variable. Standard errors are computed using cross-validation. Standard errors are in parentheses. Source: Own calculations using administrative data from the University of St. Gallen.

2.B Nonparametric estimation

The estimation of the average spillover effect, the local segregation outcome effect, and the local segregation inequality effect rely on the estimation of the mean allocation response functions, $\hat{m}_F(s)$ and $\hat{m}_M(s)$, and their derivatives, $\nabla_s \hat{m}_F(s)$ and $\nabla_s \hat{m}_M(s)$. Local private and local external peer effects follow from a decomposition of the local segregation outcome effect.

To estimate $\hat{m}_F(s)$ and $\hat{m}_M(s)$, Graham et al. (2010) propose a nonparametric estimation procedure, using kernel smoothing methods. The authors define a kernel function $K(u)$ that integrates to one (and follows other regularity conditions). Let $K_b(s - S_i) = \frac{1}{b} K\left(\frac{s-S_i}{b}\right)$. Then, the estimates of the mean allocation response function are given as

$$\hat{m}_T(s) = \frac{\hat{g}_{1T}(s)}{\hat{g}_{2T}(s)} = \frac{\frac{1}{I_T}\sum_{i=1}^{I_T} K_b(s - S_i) Y_i}{\frac{1}{I_T}\sum_{i=1}^{I_T} K_b(s - S_i)}, \tag{2.10}$$

with type $T \in M, F$. I_T denotes the number of individuals with type T in the sample, and $I = I_M + I_F$ the number of individuals in the sample. Notice that here, we abstract from conditional random assignment.

By computing the derivatives of the estimated functions, $\hat{m}_T(s)$, $\hat{g}_{1T}(s)$, and $\hat{g}_{2T}(s)$, and by plugging them into the following expression, one obtains the derivatives of the mean allocation response functions:

$$\nabla_s \hat{m}_T(s) = \frac{1}{\hat{g}_{2T}(s)} \left[\nabla_s \hat{g}_{1T}(s) - \nabla_s \hat{g}_{2T}(s) \hat{m}_T(s) \right]. \tag{2.11}$$

These objects can be plugged into the formula for the average spillover effect:

$$\hat{\beta}^{ase} = \frac{1}{I} \sum_{i=1}^{I} d_k(S_i) \{ S_i \nabla_s \hat{m}_M(S_i) + (1 - S_i) \nabla_s \hat{m}_F(S_i) \}. \tag{2.12}$$

The estimation of the local segregation outcome effect and the local segregation inequality effect follow the same estimation principle. For details and derivation of their estimation, see Graham et al. (2010).

3. Retention Effects in Higher Education

Sharon Pfister, Darjusch Tafreschi, and Petra Thiemann

Abstract

Retention policies are commonly used to maintain student quality at educational institutions. Their effectiveness, however, is debated in the literature. Existing papers investigate the effect of retention on student outcomes in primary and secondary education – results for higher education are non-existent. This paper complements the literature as it analyzes the effects of retention during the first year at the university level. To establish causality, a binding minimum requirement of the first year is utilized in a regression discontinuity framework. Administrative data from the University of St. Gallen, Switzerland, are used to estimate causal effects of retention on subsequent dropout probabilities of students, the choice of major studies, their study speed and grade performance. While the effects of retention on immediate dropout and subsequent study speed are rather modest, significant improvements in grades are found.

JEL codes: I21, I23, J24
Keywords: retention, higher education, dropout, regression discontinuity

3.1 Introduction

Student numbers in post-secondary education have increased sharply over the last decade. For example, in the US, student full-time enrollment in tertiary education has increased by 32% between 1996 and 2006 (OECD, 2008). Whereas this trend indicates an increase in educational opportunities, growing student numbers can challenge an institution's ability to maintain high educational standards as well as a high quality of the student pool, especially if universities or colleges are non-selective.

As one strategy to maintain high educational quality, many universities force or encourage underperforming students to repeat several courses or even a full year. Despite the costs that these policies impose on educational institutions, little is known about their effectiveness to persistently boost students' educational attainment.[35] Identifying a causal effect of repetition is challenging, as students are commonly selected into repetition by their universities, or self-select themselves into repeating. Thus, differences in educational outcomes between repeaters and non-repeaters might reflect differences in observable or unobservable characteristics rather than the effect of repetition.

To identify the effect of repetition, this paper exploits a strictly enforced retention policy for first-year undergraduates at the University of St. Gallen (Switzerland) in a sharp regression discontinuity design (RDD). If students do not meet a certain performance requirement (cut-off) by the end of their first year, they either have to repeat all first-year courses or have to drop out. Following a standard approach in the literature (Imbens and Lemieux, 2008, Manacorda, 2012, Jacob and Lefgren, 2004), we assume that students who are close to the cut-off and fail are comparable to students who are close to the cut-off and pass. Thus, a comparison of repeaters and non-repeaters who are close the cut-off establishes a causal effect of repeating, as long as no selective dropout occurs. Using an administrative dataset, we examine dropout, major choice, grade point averages, and credit points per semester as outcomes.

[35] Bettinger and Long (2009) find a positive impact of remedial education at the college level. Remedial education, however, refers in their paper to repetition of below-college-level courses.

The effect of grade retention on educational achievement has been extensively studied in the context of primary and secondary education, with mixed results. Theoretical predictions of the net effects of retention on individual student outcomes are also ambiguous (Manacorda, 2012). While potentially positive effects can arise because of learning gains and a better match between a students' knowledge and the level of teaching, potentially negative effects can occur because of stigmatization by both teachers and classmates, negative shocks to self-confidence, and slow adjustment to a new classroom environment. While early studies for schoolchildren emphasize the negative effects of retention (Jimerson, 2001), recent studies find rather positive effects on grades, especially for primary schoolchildren.[36] Yet, retention increases dropout of children in high school.[37] In general, the effects seem strongly age-dependent, with rather positive results in primary school, and rather negative results in high school. Results for higher education are inexistent.

Results of retention in higher education are expected to differ strongly from results of retention in primary and secondary education, for at least three reasons. First, negative effects of retention might be less pronounced in a post-secondary setting. More mature students can supposedly cope better with negative events. In addition, stigmatization by a university instructor is less likely than stigmatization by a classroom teacher, as interactions at a university take often part in larger groups on average (e.g., lectures), which diverts attention away from single students. Furthermore, detachment from initial cohort members is probably less important in a university environment, again as interactions are not restricted to a single classroom. Second, the effects of retention on dropout might be higher for university students. On the one hand, university education is voluntary, so that dropouts from university do not face any sanctions. On the other hand, outside options of university dropouts

[36] An unequivocally positive effect on test scores seems to exist for retained 3rd graders in the US. Three studies independently find a positive effect for Chicago (Jacob and Lefgren, 2004), Texas (Lorence and Dworkin, 2006), and Florida (Greene and Winters, 2007). These results, however, do not translate to 6th graders as shown by Jacob and Lefgren (2004). Roderick and Nagaoka (2005) find even negative effects on test scores of 6th graders in Chicago. All outcomes examined in these studies are short-term outcomes, that is, measured 1-3 years after retention.

[37] This result has been confirmed by Jacob and Lefgren (2009) for 6th graders in Chicago, by Ou (2010) for 9th graders in New Jersey, and by Manacorda (2012) for 7th to 9th graders in Uruguay.

are probably more valuable than outside options of high school dropouts. Third, university students might benefit especially from repeating the first year. Some students need additional time to develop new study habits, such as self-guided learning. For these students, grade repetition might provide a valuable chance of adjusting.

The paper finds overall positive effects of retention on student achievement, but also a rise in dropout in response to retention. Dropout among retained students at the cut-off is 6 percentage points higher than dropout among promoted students, but the result is not significant. Moreover, grade point averages among the repeaters in the second year are on average 0.35 standard deviations higher than grade point averages among non-repeaters. This effect persists throughout the entire observation period, that is, up to the first four semesters after repeating. The positive result on grades might come from positive selection into repeating, from knowledge improvement due to a repetition of all first-year courses, or from adjustment of study habits. Furthermore, retention affects major choice: Repeaters favor the Economics major more frequently than non-repeaters. Finally, from the second year on, repeaters display approximately the same study speed as non-repeaters in terms of credits completed per semester.

The paper proceeds as follows. Section 3.2 presents the institutional setting. Section 3.3 describes the dataset. Section 3.4 outlines the identification and estimation strategy, followed by a presentation of the results in Section 3.5. Section 3.6 concludes.

3.2 Institutional setup

The University of St. Gallen is a Swiss public institution. As a traditional business school it offers degree courses in Business Administration, Economics, International Affairs, and Legal Studies. It is the largest college of its kind at the national level when measured by the number of students in Business and Economics. The significance of the institution is reflected in Table 3.A.1, which shows the number of graduates in Switzerland and in particular at the University of St. Gallen over the last few years. The institution accounts for roughly 30% of all graduates in Economics and Business Administration in Switzerland.

For legal matters the university has no direct control about the number of new entrants. By federal law it is committed to accept every student with a Swiss university entrance license, that is, a Swiss high school certificate ("matura"). For students with foreign high school degrees there exists a pre-defined admittance rate which varies by year. Foreign students' admittance is based on an entrance test. The unrestricted admittance of Swiss high school graduates is reflected by the continuous rise in student numbers that the university experienced over the last decade. While the number of first-year students amounted to about 800 students in 2006, 10 years earlier only about 600 students entered (Table 3.A.2).

In order to maintain a high quality of education and degrees, the "assessment year" (ASY) was introduced in 2001. The primary goal of the ASY is to select first-year students into the Bachelor level (second year). Students are allowed to proceed to the Bachelor level when they meet the requirements as stated by the ASY regulations. Over the years 2001–2006, the university admitted only approximately two thirds of all first-year students to the Bachelor directly. Non-admittance can be due to both voluntary and non-voluntary dropout.

The ASY requires identical core subjects and test criteria for all students. By the end of the ASY, students must have chosen their Bachelor specialization (Business, Economics, International Affairs, Law and Economics. There are two subgroups of students for which the ASY differs. First, students who intend to specialize in legal studies follow a different curriculum during the ASY. Second, students of non-German mother tongue can choose to complete the assessment courses within two years instead of one ("extended track"). Because of their special status, both groups are excluded from all analyses in this paper. All other students follow a strictly defined standard curriculum (Table 3.A.3), henceforth denoted as "Business/Economics track", and form the population of interest for this study.

The core curriculum of the Business/Economics track consists of courses in *Business Administration, Economics, Legal Studies, and Mathematics*. These subjects are tested after the first and second semester during predetermined examination weeks (“central examination periods”). Moreover, students have to proof sufficient foreign language skills, which are also examined during the second central examination period. In addition to the core subjects, students have to take courses in *leadership skills* and *critical thinking*. Both are either evaluated on the basis of group presentations or written essays. Finally, students have to submit a major essay in one of the core subjects to be handed in before registering to the second central examination period.

All courses are compulsory, and each course in the ASY is graded and weighted by a number of predefined credit points. The overall grading of all courses, in turn, leads to the final decision on whether the student passes or fails the ASY. The whole curriculum of the ASY consists of a total of 60 credits (55 in 2001). The grading scale in the Swiss education system is defined from 1 to 6 in steps of 0.5, with 4 being the worst passing grade and increasing values indicating better performance. The grading process makes sure that students cannot pass or fail only due to one evaluation. Instead, a student passes or fails according to a *performance measure based on all course grades*. Students who do not pass have a one-time opportunity to repeat the first year (“retention treatment”).

The ASY is designed so as to make selection into the retention treatment as objective and non-manipulable as possible. This is ensured by the following four steps. First, all courses are compulsory, and examination dates are fixed by the university. Second, examination dates are blocked in short time periods. In particular, all core subjects are tested within a central examination period of five weeks in both the fall and the spring term. This time pattern leaves only limited space for students to strategically adjust their learning behavior during these periods. Third, grade disclosure takes place exclusively at the end of each semester. All course grades are jointly disclosed on the same day by mail. Notice that exams taken in the final central examination period take place in calendar week 25–29 and account for 25.5 credits (Table 3.A.3). Students receive no information on their performance in these courses before calendar week 35. Thus, students receive no information about pivotal grades during this examination period and therefore there exists no meaningful strategic behavior when exams are taken. Fourth, students have hardly any chance to enforce a revision of grades.[38]

Once all compulsory courses are completed, the ultimate criterion for passing the ASY is threefold. First, individuals have to accumulate 240 (220 in 2001) credit-grade-points, which corresponds to an overall average grade of 4.0. Second, individuals must not accumulate more than 12 "minus credits" (MC), a rule that we will elaborated on below. Third, individuals have to submit a proof of sufficient accounting skills.

The further analysis will concentrate on the second criterion, for the following reasons. The first criterion is rarely violated if the other criteria are fulfilled, that is, only 3 individuals in the cohorts of 2001-2006 failed the ASY because of an insufficient number of credit-grade-points. These individuals are excluded from the analysis. Furthermore, violating the third criterion implies neither passing nor failing the ASY. Once the other two criteria are fulfilled, violating the third criterion allows for conditional acceptance into the Bachelor. Yet, students have to pause and submit a proof of accounting skills in the meantime. As this group is particular and rather small, we will also exclude them from the analysis (see section 3.3).

Thus, the decisive rule for passing versus retention is given by a strict threshold of 12 minus credits which will later be exploited as a treatment rule. For any course, a student receives minus credits if he obtains a grade below 4. Minus credits in the respective course are then defined as the difference between the actual grade and a grade of 4, multiplied by the number of credits for this course. For example, if a students receives a grade of 3.5 in a course with 4 credits, he obtains 2 minus credits for this course. To describe the accumulation of minus credits more formally, suppose that the overall number of compulsory subjects in the ASY is S. Let G_s be the grade obtained in course s, and C_s the number of credit points associated with

[38] Only in the case of obvious mistakes during grading, grade revision is unequivocally granted. In all other cases, the individuals have to file a case ("recourse"). In the data for 2001-2006, we observe 2 cases with insufficient performance according to the data who are still observed in the Bachelor afterwards. These individuals might have won a case for grade revision.

this course. The total sum of minus credits (MC) is then calculated as

$$MC = \sum_{s=1}^{S}(4 - G_s) * C_s * \mathbf{1}(G_s < 4) \tag{3.1}$$

Note that minus credits cannot be compensated for by grades greater than 4 in other subjects. If $MC > 12$, the respective student fails the ASY and cannot directly proceed to the Bachelor level, that is, the student is retained. Yet, he is allowed to repeat the *full* ASY. If the ASY is successfully passed in the second attempt, the student is admitted to the Bachelor level and can proceed (see Figure 3.A.1). In case of failing again, the student is coercively exmatriculated. Enforcement of this rule is strict.

Prior to entering the Bachelor level, individuals choose their Bachelor specialization (Business, Economics, Law and Economics, or International Affairs). During the Bachelor phase, they complete a number of compulsory courses and electives as well as a Bachelor thesis, but are free to set the pace of degree completion themselves. On-time graduation follows after 4 Bachelor semesters, but only a minor fraction of students manages to complete their Bachelor degree within this time frame. Yet, conditional on entering the Bachelor level, more than 84% of students graduate within at most 6 Bachelor semesters.

3.3 Data

3.3.1 General description

The analyses in this paper are based on administrative records from the University of St. Gallen. The data consist of enrollment and course data at the student level and cover the population of all students entering the ASY between 2001 and 2006. Students entering the ASY after 2006 are excluded from the analysis as long-term outcomes are unobserved for this group (i.e., for the latest cohorts we only have incomplete information about their performance at the Bachelor level).

The enrollment data contain the following information. First, we can observe whether the ASY has been passed successfully or not. Second, for each student who fails the first year, we observe whether he repeats the ASY or drops out, respectively. Third, the data contain information about major choice at the Bachelor level. Fourth, we observe individual characteristics at the date of university entry, that is, age, sex, nationality, mother tongue, country of origin, as well as region of origin for students from Switzerland, type of high school degree and in which country or region it has been obtained, as well as the date of high school graduation.

The course file contains information on individual performance at the course level for each semester, that is, grades and credits for each completed course. This information is crucial at different points in the analysis. First, on the basis of course information, we restrict the estimation sample to first-year students who have completed all compulsory courses as required by the curriculum. Second, the course file allows us to compute the precise number of minus credits obtained by each student. Third, we use grades and credit points at the Bachelor level as measures of academic performance. In order to capture the pace of degree completion, we use credits obtained by the end of each Bachelor semester. As a measure of the quality of performance, we use standardized grades. Grades are standardized at the level of the Bachelor entry cohort.

The initial sample consists of all entering first-year students with German mother tongue who start the Business/Economics track at the University of St. Gallen (n = 3,762, see Table 3.A.4). This sample is homogeneous in the sense that first, all Business/Economics students have to complete the same courses during their freshman year, and second, these students face exactly the same exam conditions. The latter is not the case for students with foreign mother tongue as they might have longer exam durations. Table 3.A.4 shows that the composition of the student cohorts in terms of background characteristics remains approximately stable over the years. The large majority of students is male (73%). Moreover, most students enter the ASY when they are 20 or 21 years old. Foreign students account for 22% of students on average. This number is low due to the admission rules: The fraction of students with neither high school diploma from Switzerland nor Swiss citizenship is restricted to at most 20% of the student body.

For the main analyses in this paper we only consider students who have completed all first-year courses. For all other freshman students, the assumption of random assignment to retention is problematic, since course non-completion might relate to unobserved characteristics which also affect subsequent outcomes. Therefore, the following types of students are dropped: (i) students who have not completed all mandatory first semester courses (type 1), (ii) students who have completed all mandatory first semester courses, but have exceeded the threshold of 12 MC already in the first semester (type 2), (iii) students who have passed the first semester but have dropped out voluntarily after the first semester (type 3), (iv) students who have not completed all courses in their second semester (type 4), (v) students who fail the mandatory accounting test and therefore cannot be promoted (type 5). All other students (type 6) are included in the final estimation sample (n = 2,983). The estimation sample is therefore a selected sample of all entering first-year students. We describe the different types of students in Section 3.3.2 to deepen our understanding about the selection process throughout the first year.

3.3.2 What happens during the first year?

Our final estimation sample contains only students who complete all first-year exams (see Table 3.1). These account for 79% of all freshmen with German mother tongue who enter the Business/Economics track. Accordingly, 21% of all freshmen miss at least one requirement of the ASY. To understand the selection dynamics throughout the first year, we examine several pathways that lead to the exclusion from our final estimation sample. For this purpose, we classify students according to 6 types as mentioned in Section 3.3.1. Classification is based on three criteria: first, whether all main exams have been taken (first semester: Business 1, Econ 1, Math 1, second semester: Business 2, Econ 2, Math 2, Term paper), second, whether students have exceeded the threshold of 12 MC in the first or second semester, respectively, and third, the point in time when they drop out. An additional criterion is passing the mandatory accounting exam by the end of the first year. Students who have not passed this exam are "blocked", that is, they must not take any courses for at least one semester. Table 3.1 shows the relative size for each of the type subgroups.

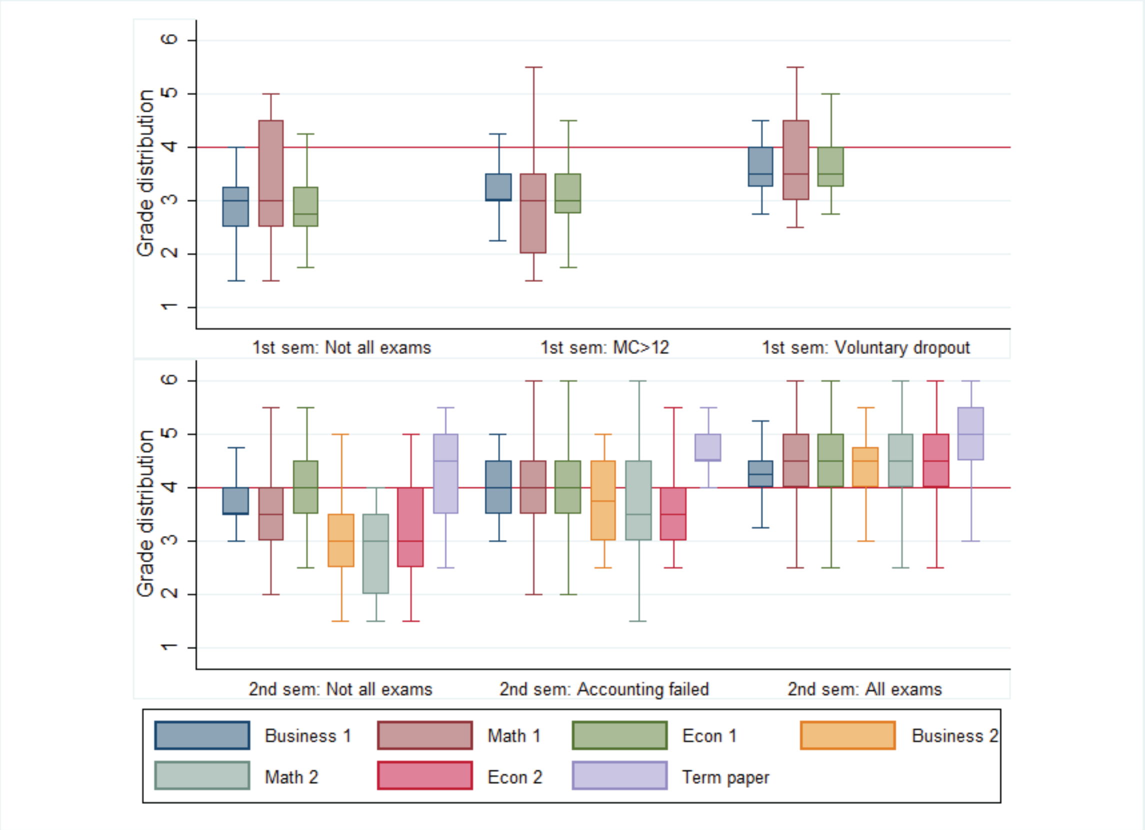
Table 3.1: Description of student types

Type	Main exams taken	MC 1st sem.	Dropout	# Obs.	% of sample
1	Max. 2 in 1st sem.	-	Voluntary dropout in 1st sem.	225	6%
2	All in 1st sem.	> 12	Blocked for 2nd sem.	365	10%
3	All in 1st sem.	≤ 12	Voluntary dropout after 1st sem.	26	1%
4	All in 1st, but none in 2nd sem.	≤ 12	Voluntary dropout in 2nd sem.	90	2%
5	All in both sem.	≤ 12	Delay (missing accounting exam)	73	2%
6	All in both sem.	≤ 12	No dropout / Voluntary dropout	2,983	79%
Total				3,762	100%

The sample includes all first-year students with German mother tongue entering in 2001–2006 into the Business/Economics track. Type definitions are presented in Section 3.3.3.

Dropout is particularly high at the beginning of the first year. Among the first-semester dropouts, we can distinguish between three groups. First, 6% of students do not complete all main first-semester exams (type 1). These students might have been discouraged already in the beginning. Second, 10% of students take all required exams, but exceed the threshold of 12 MC already by the end of the first semester (type 2). As a result these students are not allowed to enter the second semester. Third, a minor fraction of 1% decides to dropout despite successfully passing his first-semester courses (type 3). Among the three groups, type-1-students appear as the lowest performing students (see Figure 3.1). Considering only the exams that these students have taken, their median performance is slightly lower than the performance of the median type-2-student. Unsurprisingly, late voluntary dropouts (type 3) have better grades than early voluntary dropouts (type 1). Yet, the performance of a type-3-student is often below the passing grade of 4. Overall, a visible descriptive relationship between performance and dropout exists during the first semester. It is however unclear whether students drop out due to their low expected performance, which would be in line with the idea of "schooling as experimentation" (Manski, 1989), or whether dropout and performance are confounded by other unobserved factors, or both.

Figure 3.1: Grades by type



Box-plots of the grade distributions, by type. The sample includes all first-year students with German mother tongue entering in 2001 - 2006 into the Business/Economics track. Type definitions are presented in Section 3.3.3.

Dropout during the second semester is low, that is, only 4% of students drop out after having entered the second semester. Half of them drop out voluntarily as they do not take all required exams in the second semester (type 4), and the other half is delayed due to a missing accounting exam. Again, better performing students tend to drop out at a later stage or to stay (Figure 3.1).

Table 3.2 also gives some indication about the extent to which types might be confounded by observed characteristics. With respect to age, no clear pattern exists. By contrast, foreign students and students with an entrance test tend to stay longer. This observation is also in line with the considerations of Manski (1989) as initial costs of taking up studies in St. Gallen are higher for foreign students (e.g. studying for the entrance exam, moving to a foreign country, higher student fees for foreign students), and thus their initial performance expectations must also be higher in order to have positive expected gains of taking up a degree. Similarly, students who completed their high school diploma in the canton of St. Gallen are overrepresented among the first three types as they might have had lower costs in the beginning.

Table 3.2: Descriptive statistics: Entering first-year students by type

	(1)	(2)	(3)	(4)	(5)	(6)
	1st sem: Not all exams	1st sem: MC > 12	1st sem: Voluntary dropout	2nd sem: Not all exams	2nd sem: Accounting failed	2nd sem: All exams
Background Characteristics						
Age $\leq$ 19	10%	15%	12%	17%	15%	12%
Age 20/21	52%	53%	62%	51%	56%	68%
Age $\geq$ 22	38%	32%	27%	32%	29%	21%
Nationality foreign	20%	16%	0%	19%	37%	23%
Entrance test	5%	5%	0%	12%	33%	18%
Matura in SG	20%	16%	12%	14%	23%	15%
Minus Credits						
# MC in 1st sem.	-	19.09	8.38	7.53	4.05	1.60
# MC in first year	-	-	-	-	13.85	4.26
Failed: MC > 12	-	-	-	-	53%	11%
Course Performance						
Business 1 taken	28%	-	-	-	-	-
Grade Business 1	2.90	3.09	3.60	3.71	3.96	4.24
Math 1 taken	5%	-	-	-	-	-
Grade Math 1	3.27	2.83	3.69	3.58	3.99	4.54
Econ 1 taken	26%	-	-	-	-	-
Grade Econ 1	2.98	3.15	3.68	3.91	4.15	4.60
Business 2 taken	-	-	-	34%	-	-
Grade Business 2	-	-	-	3.07	3.77	4.34
Math 2 taken	-	-	-	26%	-	-
Grade Math 2	-	-	-	2.76	3.86	4.56
Econ 2 taken	-	-	-	34%	-	-
Grade Econ 2	-	-	-	3.27	3.76	4.46
Term paper taken	-	-	-	64%	-	-
Grade Term paper	-	-	-	4.36	4.77	4.97
Bachelor						
Bachelor started	10%	26%	-	43%	56%	97%
# obs	225	365	26	90	73	2,983

The sample includes all first-year students with German mother tongue entering between 2001 and 2006 into the Business/Economics track.

3.3.3 Estimation sample

Table 3.3 describes the final estimation sample (type 6). Cohort characteristics are relatively stable over the years, except for slight changes in the age distribution. Moreover, the estimation sample is comparable to the initial overall sample of entering first-year students in terms of observable student background characteristics (see Tables 3.A.4 and 3.3). Nevertheless, the sample might differ from the overall freshman population in terms of unobservable characteristics such as motivation or ability (see Section 3.3.2).

Regarding the educational outcomes of the estimation sample we observe the following: First, the fraction of students who do not accumulate any minus credits amounts to 56% on average. Moreover, students tend to accumulate larger amounts of minus credits in the second semester. In total, 11% of students fail the ASY, and 10% of students are repeaters, implying that dropout rates after retention are low. Moreover, 97% of students in the estimation sample are observed to enter the Bachelor level at some point. This number implies that most of the repeaters must have repeated successfully.

Treatment assignment is as good as random in the specified sample at the cut-off of 12 minus credits as argued in Section 3.2. Table 3.4 descriptively supports this point. First, mean background characteristics are approximately stable at the cut-off. Age is slightly lower at the cut-off, but on both sides of the cut-off, which does not threaten the analysis. The most worrisome observation is that the fraction of individuals with an entrance degree from St. Gallen is 10 percentage points higher among students just below the cut-off. Second, mean grades in the main subjects are also smoothly distributed. Given that grades are smoothly distributed, manipulation on the part of graders seems unlikely. Third, in terms of outcomes, dropout rates after the first year increase sharply at the threshold but are still low for retained students. This fact is also reflected by the high number of students starting the Bachelor degree to both sides of the cutoff.

Table 3.3: Descriptive statistics: Estimation sample by year

	Obs.	2001-6	2001	2002	2003	2004	2005	2006
Characteristics								
Male	2,983	74%	74%	74%	75%	75%	75%	72%
Age < 20	2,983	12%	8%	8%	13%	15%	12%	17%
Age 20/21	2,983	68%	68%	67%	69%	69%	70%	63%
Age > 21	2,983	21%	24%	24%	18%	16%	18%	20%
Foreign nationality	2,983	23%	19%	24%	25%	27%	24%	21%
High school St. Gallen	2,983	15%	13%	14%	16%	18%	14%	16%
Entrancetest	2,983	18%	15%	18%	21%	21%	17%	16%
Grades								
Business 1	2,983	4.24	4.24	4.24	4.20	4.23	4.26	4.28
Math 1	2,983	4.54	4.43	4.62	4.49	4.64	4.55	4.56
Econ 1	2,983	4.60	4.60	4.50	4.56	4.81	4.63	4.55
Business 2	2,983	4.34	4.32	4.14	4.46	4.79	4.04	4.44
Math 2	2,983	4.56	4.53	4.47	4.61	4.89	4.56	4.40
Econ 2	2,983	4.46	4.41	4.30	4.26	4.44	4.54	4.77
Term paper	2,983	4.97	4.90	5.00	4.92	4.98	5.00	5.03
Minus Credits								
MC > 0 in first year	2,983	56%	63%	59%	56%	51%	55%	50%
# MC in first semester	2,983	1.60	1.51	1.54	1.85	1.47	1.69	1.56
# MC in first year	2,983	4.26	4.16	4.88	4.14	3.78	4.94	3.62
Fail Total	2,983	11%	12%	12%	10%	9%	15%	9%
Retention								
Repeater	2,983	10%	10%	11%	9%	9%	14%	8%
Bachelor								
Bachelor started	2,983	97%	95%	96%	98%	98%	96%	97%
Bachelor $\leq$ 4 semesters	2,886	37%	47%	50%	39%	33%	26%	23%
Bachelor $\leq$ 5 semesters	2,886	61%	69%	71%	64%	60%	55%	49%
Bachelor $\leq$ 6 semesters	2,886	84%	87%	89%	88%	86%	80%	77%
Observations		2,983	609	498	426	380	511	559

The sample includes all first-year students with German mother tongue entering in 2001–2006 into the Business/Economics track who have completed all compulsory courses. The last three rows (Bachelor duration $\leq$ 4/5/6 semesters) are only specified for students starting a Bachelor degree.

Table 3.4: Descriptive statistics: Estimation sample by minus credits

	Obs.	Total	0-4	4-8	8-10	10-12	12-14	14-16	>16
Background Characteristics									
Male	2,983	74%	74%	73%	75%	73%	71%	73%	75%
Age < 20	2,983	12%	13%	10%	17%	6%	9%	5%	13%
Age 20/21	2,983	68%	70%	65%	56%	69%	63%	58%	63%
Age > 21	2,983	21%	18%	25%	26%	24%	29%	38%	24%
Foreign nationality	2,983	23%	26%	16%	17%	15%	14%	13%	18%
Entrance degree from SG	2,983	15%	14%	16%	10%	23%	13%	17%	18%
Entrancetest	2,983	18%	23%	10%	10%	7%	4%	6%	7%
Grades									
Business 1	2,983	4.24	4.44	3.99	3.96	3.76	3.70	3.73	3.64
Math 1	2,983	4.54	4.83	4.20	4.00	3.92	3.89	3.76	3.61
Econ 1	2,983	4.60	4.88	4.24	4.04	3.95	3.89	3.88	3.82
Business 2	2,983	4.34	4.58	4.10	4.03	3.92	3.71	3.64	3.42
Math 2	2,983	4.56	4.89	4.17	3.93	3.97	3.95	3.67	3.39
Econ 2	2,983	4.46	4.77	4.17	3.94	3.80	3.70	3.64	3.33
Term paper	2,983	4.97	5.10	4.83	4.71	4.70	4.75	4.52	4.56
Minus Credits									
# MC in first semester	2,983	1.60	0.27	2.37	3.52	4.59	5.35	5.41	7.38
# MC in first year	2,983	4.26	0.69	6.00	9.09	11.02	13.08	15.05	21.78
Retention and Dropout									
Repeater	2,983	10%	-	-	-	-	91%	92%	88%
Dropout after 2 sem	2,983	2%	1%	0%	0%	2%	9%	8%	13%
Bachelor									
Bachelor started	2,983	97%	99%	99%	100%	98%	83%	83%	74%
Bachelor $\leq$ 4 semesters	2,886	37%	41%	31%	27%	21%	21%	25%	22%
Bachelor $\leq$ 5 semesters	2,886	61%	68%	52%	45%	37%	45%	49%	44%
Bachelor $\leq$ 6 semesters	2,886	84%	90%	76%	73%	74%	66%	70%	63%
Observations		2,983	1'975	432	126	108	70	64	208

The sample includes all first-year students with German mother tongue entering in 2001 - 2006 into the Business/Economics track who have completed all compulsory courses. The last three rows (Bachelor duration $\leq 4/5/6$ semesters) are only specified for students starting a Bachelor degree.

3.4 Identification strategy and estimation

The passing requirements for first-year students provide a strictly enforced rule that allows us to obtain a local estimate of the causal effect of retention and repetition on future educational outcomes, that is, we use the threshold value (c) of 12 minus credits for identification. We argue in the following that retention is quasi-random at the threshold.

In the potential outcome framework, Y_i^0 and Y_i^1 are the outcomes of individual i in a state without and with retention, respectively. For each student, only one state is observed at any moment in time, that is,

$$Y_i = Y_i^1 * R_i + Y_i^0 * (1 - R_i), \tag{3.2}$$

with $R_i = 1$ if the individual is retained, and $R_i = 0$ if the individual is promoted. We are interested in the average treatment effect of retention on future outcomes,

$$\begin{aligned} \tau &= E[Y^1 - Y^0] && (3.3) \\ &= \pi * E[Y^1 - Y^0 | R = 1] + (1 - \pi) * E[Y^1 - Y^0 | R = 0], && (3.4) \end{aligned}$$

where Y is the educational outcome of interest, R is the binary retention (treatment) status, which jumps from zero to one at the cut-off value of 12 minus credits, and π is the fraction of retained students.

However, the average treatment effect cannot be revealed from the data without further assumptions, because the retention status is non-randomly assigned. A conditional mean comparison (conditional only on observable characteristics) between all students for which $R = 1$ and all students for which $R = 0$ would reveal the treatment effect of interest only if unobservable characteristics were identically distributed in both groups. Since, however, unobservable characteristics (e.g., motivation or ambition) might be systematically different across the two groups (e.g., students that pass are more ambitious), a conditional mean comparison will lead to a biased estimate of the retention effect.

Yet, the group of students who *just* passed is expected to be fairly similar in terms of their distribution of unobservable characteristics, compared to the group of students who *just* failed the first year. Consequently, being retained is assumed to be quasi-random around the threshold. Following this logic, we restrict the identification of the retention effect to this local sub-population.

In the related literature, this assumption is known as the *local continuity assumption* (Imbens and Lemieux, 2008). The assumption implies the following: If the individuals within a small window around the threshold were exposed to the same policy, they would have achieved on average the same outcome. Hence, as a consequence of a student's inability to precisely control the number of minus credits (MC) achieved, retention is as good as random in the local neighborhood around the threshold and thus allows us to identify the effect at the discontinuity point c, that is,

$$\tau_{RD} = E[Y^1 - Y^0 | MC = c]. \tag{3.5}$$

While we can test whether the observed covariates X are similarly distributed around the threshold, the assumption that this is also true for unobserved characteristics is an identifying assumption that cannot be tested.

The *local continuity assumption* further implies that, for our identification strategy to be valid, we have to ensure that students who end up close to the critical threshold could not perfectly anticipate on which side of the threshold they will be placed. This assumption seems reasonable in our setting. While students are aware of the threshold *ex-ante* (as it is announced in the rules of the ASY) it is unlikely that they are able to sort themselves just above or just below the threshold *once they have taken all exams.* Another argument against strategic sorting is that grades and minus credits are not perfectly predictable from the perspective of the student. Grading schemes are often designed after the exams are taken and the students do not have any control about that process. Moreover, grading schemes are solely decided upon by the teachers.[39]

[39] Hence, the only way to purposely achieve a position just above the threshold is to apply for a revision of grades. However, from administrative sources we know that the number of individuals

In addition to these arguments, we examine the assumption of *local continuity* by investigation of the density of minus credits achieved on either side of the cut-off value, as proposed by Lee and Lemieux (2010). If density plots are smooth at the threshold, we can be confident that no sorting took place.

Continuity is depicted by Figure 3.2, which shows a histogram of the (recentered) number of minus credits. The aggregated bins depict a smoothed version of the distribution of minus credits. The vertical line represents the cut-off value. No visual indication of sorting around the threshold exists. Likewise, the McCrary test (McCrary, 2008) does not reject the null-hypothesis of continuity of the running variable at the cut-off c (log difference in height = 0.05, p-value = 0.75).

Further using the threshold value in our identification strategy, we examine the effects of retention and repetition on various outcomes: dropout probability after the first year (binary), whether a student is ever observed at the Bachelor level (binary), choice of major studies (binary) as well as continuous educational outcomes (credits and grades) over the subsequent semesters at the Bachelor level.

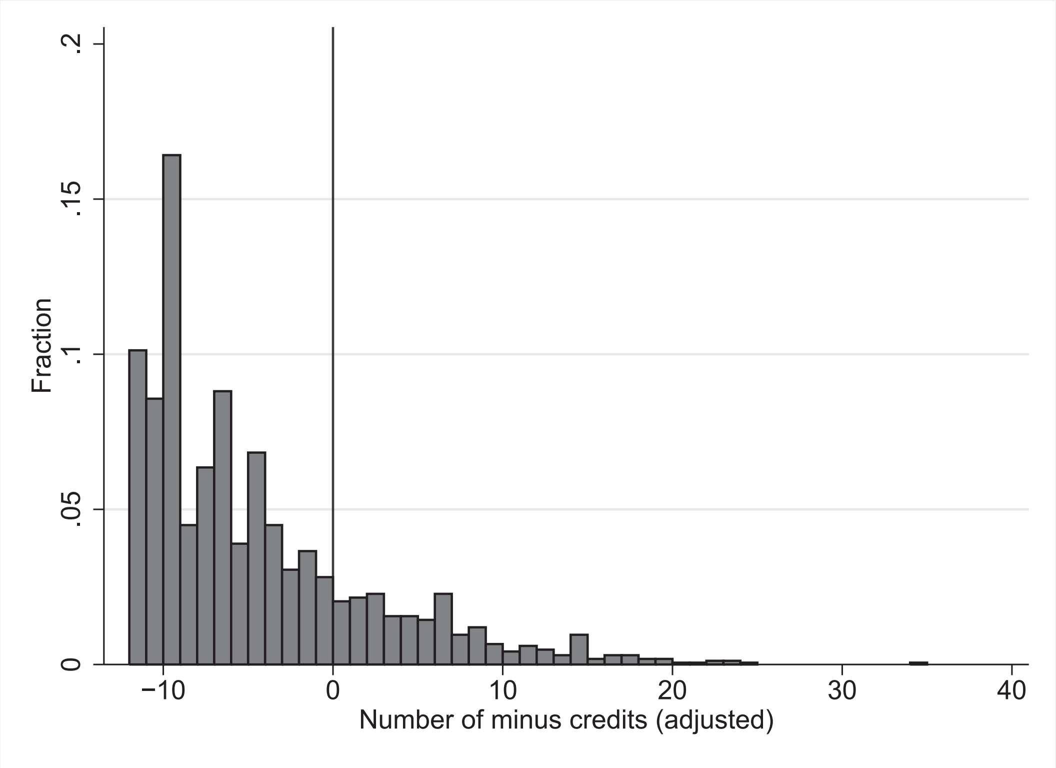
In addition to simple mean outcome comparisons at the threshold, we estimate models of the following form,

$$Y = \alpha + \beta * 1(MC \geq 0) + \sum_{k=1}^{K} \gamma_k * MC^k + \sum_{k=1}^{K} \nu_k * MC^k * 1(MC \geq 0) + \varepsilon, \quad (3.6)$$

where Y represents the educational outcome (e.g., grade point average), MC corresponds to the recentered (minus 12.25) number of minus credits collected by the student at the end of the first year, k is a flexibility parameter, and ϵ is an error term. In all specifications, the coefficient of interest is β, representing the causal effect of retention on the outcomes. We furthermore use varying windows of data around the threshold to assess the robustness of our findings. By using higher order polynomials and interaction terms, we allow for a non-linear relationship as well as different slopes on both sides of the cut-off. We also provide nonparametric

who manage to shift themselves below the critical cut-off value as a result of a revision process is, if anything, marginal (in most years even non-existent), as explained in Section 3.2.

Figure 3.2: Histogram of the assignment variable



The assignment variable is defined as the amount of minus credits accumulated during the ASY. Minus credits are adjusted by subtracting the cut-off value (12.25 minus credits) from the actual amount of minus credits. The sample consists of all individuals in the estimation sample (cohorts 2001–2006) who have accumulated at least 0.25 minus credits (n = 1,669).

estimates based on the guidelines provided by Imbens and Lemieux (2008). These are based on kernel methods (where the optimal bandwidth is computed by cross-validation) using local linear regressions to estimate the boundary points on each side of the threshold. As with our parametric specifications, the effects of interest are identified by the differences in the expected means of the outcomes on either side of the threshold. The results for our preferred specifications are reported in the result section, while various other specifications are to be found in the appendix.

Using the same model specifications, we also investigate the local continuity of predetermined covariates around the threshold to address concerns related to strategic sorting. Table 3.5 presents the coefficient estimates. While there is no evidence for gender or age-related sorting around the threshold, there is some indication for discontinuities with respect to the origin of students, that is, we find some evidence that students who just fail are less likely to be from "nearby St. Gallen" (as already mentioned in the data section). To account for such differences, we estimate all models for educational outcomes with and without covariates (covariates included correspond to the ones in Table 3.5) to check the robustness of our findings.

Our identification strategy faces two further difficulties. First, we are interested in performing same-grade comparisons, that is, educational outcomes of retained students are compared to the same outcomes of non-retained students. By the nature of the problem, however, the outcomes of the two groups are measured one year apart (the outcome of retained students is typically measured one year later). However, it might well be that outcome distributions are not *per se* comparable across years. To account for potential problems that are due to changes in grade distributions over time, we standardize on the level of the Bachelor level entry cohort.

Second, the problem of non-random dropout might bias the results when we investigate future educational outcomes other than dropout decisions that occur right after the first year. As discussed, retention might induce individuals to leave university (instead of repeating the first year). The identification results presented so far are only valid under the assumption of random dropout. If dropout were selective, our estimates would be biased. Assume that the less able students are more likely to drop out as a result of failing the first year, that is, only the more able remain and repeat. In such a setting, one might estimate a positive effect of repeating the first year on subsequent outcomes (while the true effect is zero) just because of selective dropout. We clearly acknowledge such concerns. Yet, out of the 342 students in our estimation sample who are retained only 37 students (or 11%) drop out. Moreover, this number is smaller at the threshold (around 7%). We do not consider this share substantial and thus abstain from extending the analyses towards partial identification strategies that would only allow to estimate bounds for the effects of interest.

3.5 Results

This section provides visual evidence as well as regression estimates of the effects of retention on academic outcomes based on the regression discontinuity design (i.e., for students close to the cut-off value of 12 minus credits). First, we investigate how retention affects the dropout decisions of students after their first year, that is, before enrollment into their third semester. To which extent does retention cause students to drop out, immediately or throughout their second attempt of the ASY, respectively? Second, we examine the effect of retention on major choice. Finally, we compare the academic outcomes of repeaters and non-repeaters at the Bachelor level. In particular, we focus on the number of credits accumulated in subsequent semesters as well as the corresponding grade point averages (GPA) by the end of each Bachelor semester.

3.5.1 The effect of retention on dropout

Interpreting dropout decisions as a utility maximization problem, Manski (1989) points out that students weigh their expected utility from dropping out against their continuation utility. Within the institutional framework we explained earlier, retained students incur higher continuation costs than non-retained students. They face not only one additional year of foregone earnings, but also the risk of failing the ASY for a second time. In addition, being separated from their entry cohort or being stigmatized as a repeater might be associated with additional (psychological) costs.

Table 3.5: RDD estimates: Pre-determined characteristics

Dependent variable	(1)	(2)	(3)	(4)
Male	-0.04	-0.14	-0.17	-0.05
	(0.09)	(0.10)	(0.15)	(0.14)
Age 17/19	0.02	0.04	-0.04	-0.02
	(0.10)	(0.11)	(0.15)	(0.15)
Age 20/21	-0.04	-0.06	-0.10	-0.07
	(0.10)	(0.11)	(0.16)	(0.17)
Age 22+	0.02	0.01	0.14	0.05
	(0.05)	(0.06)	(0.09)	(0.09)
Gap year	-0.06	-0.10	-0.14	0.02
	(0.10)	(0.12)	(0.16)	(0.17)
High school St. Gallen	-0.13	-0.15*	-0.21*	-0.24*
	(0.08)	(0.09)	(0.13)	(0.14)
Foreign citizen	0.00	0.01	0.04	-0.02
	(0.07)	(0.08)	(0.11)	(0.11)
Entrance test	-0.02	0.00	-0.01	-0.03
	(0.05)	(0.05)	(0.07)	(0.06)
Estimation Window	[1;1]	[8;8]	[12;12]	[12;12]
Polynomial order	0	2	4	NP
Observations	94	908	1615	1615

RDD estimates of dropout behavior and major choice. Columns (1)-(4) display different specifications. The parametric specifications (1)-(3) are estimated using a linear probability model. Following Imbens and Lemieux (2008) the bandwidth for the local-linear nonparametric specification (NP) is determined by cross-validation. The respective estimation window for each specification is reported as the minus credit range on each side of the threshold.
* Significant at 10%- level, ** Significant at 5%- level, *** Significant at 1%- level. Standard errors in parentheses.

These considerations suggest that dropout rates are supposedly higher for retained students.

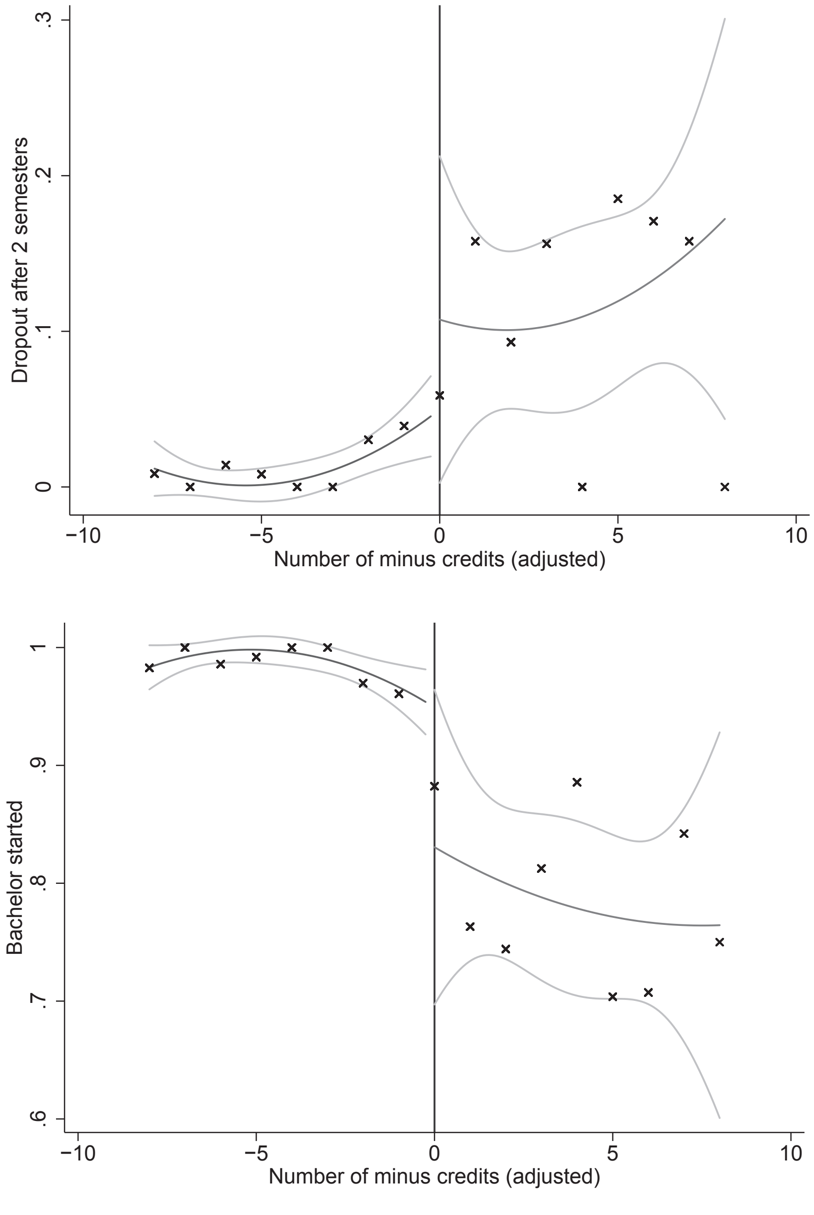
Figure 3.3 shows students' mean dropout rates (left panel) as well as their probability to enter the Bachelor level (right panel) as a function of their minus credits by the end of their first year. The threshold value is depicted by the vertical line, together with quadratic regression lines (fitted on each side separately). The vast majority ($> 97\%$) of students who meet the passing requirements of the first year do not drop out and proceed with their Bachelor studies. On the contrary, the share of retained students who drop out immediately after their first year is larger (approximately 7% for students close to the threshold value). The probability to be observed at the Bachelor level at some point is also lower for retained students. The figure indicates that of those students who just failed their first attempt, 87% are observed at the Bachelor level later on. Given the large variances in both outcomes for the sub-sample of retained students, we can expect regression estimates to depend on the flexibility of the underlying model as well as the chosen window that they are based upon.

Table 3.6 presents various regression estimates of the effects of retention on dropout immediately after the first year.[40] We run numerous specifications of the model, but only show the results based on our preferred model specifications. In order to account for the trade-off between bias and precision, the specifications grow more flexible as the sample size (i.e., the estimation window) increases (Lee and Lemieux, 2010). The standard errors of the coefficient estimates eventually increase – despite the larger bandwidths – as a result of the higher order polynomials. We also enrich the specification to include covariates in order to balance potential differences that are due to observable characteristics. In addition to parametric models, we provide non-parametric estimates using the approach suggested by Imbens and Lemieux (2008).[41]

[40] Dropouts are defined as students who are not observed to enroll for the third semester.

[41] We only consider results based on samples that exclude students with zero minus credits, as they are uninformative for our purposes.

Figure 3.3: RDD estimates: Probability of immediate dropout and starting a Bachelor degree



The panels above provide a graphical illustration of the RDD estimates of the probabilities of dropout after 2 semesters and starting a Bachelor degree. The green dots represent the mean outcomes within each minus credit category. The green lines display a quadratic fit to either side of the cutoff (95% confidence intervals in gray). The sample consists of all individuals in the sample within a range of 8 minus credits to either side of the cutoff (n = 908).

Table 3.6: RDD estimates: Dropout and major choice

Outcome	(1)	(2)	(3)	(4)	(5)
			Selection		
Dropout after	0.06	0.04	0.00	0.01	-0.01
2nd semester	(0.05)	(0.05)	(0.06)	(0.06)	(0.10)
Bachelor Started	-0.15***	-0.12**	-0.04	-0.05	-0.05
	(0.06)	(0.06)	(0.08)	(0.08)	(0.11)
Estimation Windows	[1;1]	[8;8]	[12;12]	[12;12]	[12;12]
Polynomial order	0	2	4	4	NP
Observations	94	908	1,615	1,615	1,615
Covariates	No	No	No	Yes	No
			Major Choice		
Economics major	0.10*	0.18***	0.09	0.11	0.10*
	(0.05)	(0.06)	(0.08)	(0.08)	(0.07)
Business major	-0.12	-0.22	-0.20	-0.21	-0.16
	(0.10)	(0.11)	(0.14)	(0.14)	(0.13)
Estimation Window	[1;1]	[8;8]	[12;12]	[12;12]	[12;12]
Polynomial order	0	2	4	4	NP
Observations	82	815	1484	1484	1484
Covariates	No	No	No	Yes	No

The table shows RDD estimates of the the effect of student retention on dropout behavior and major choice. Columns (1)-(5) display different specifications. The parametric specifications (1)-(4) are estimated using a linear probability model. Following Imbens and Lemieux (2008) the bandwidth for the local-linear nonparametric specification (NP) is determined by cross-validation. The respective estimation window for each specification is reported as the minus credit range on each side of the threshold. Covariates include the following indicator variables: Cohort dummies, Male, Younger than 20 by the start of the ASY, Older than 21 by the start of the ASY, Non-Swiss nationality, Non-German mother tongue, High school St. Gallen, Entrance test participation, Gap year after finishing high school.

* Significant at 10%-level, ** Significant at 5%-level, *** Significant at 1%-level. Standard errors in parentheses.

In line with the visual evidence, most (i.e., all parametric) point estimates are positive. Moreover, standard errors generally increase despite growing bandwidths as a consequence of the added model flexibility. The non-parametric estimate in column (5) is negative, but also suffers from a larger standard error.[42] Overall, none of the estimates with respect to immediate student dropout is statistically significant, which is supposedly caused by the relatively large variance in dropout rates for students above the cut-off point. The largest point estimate is found by the simple mean comparison in column (1), which shows a 6 percentage points higher dropout rate for the retained students, which we still consider modest. Hence, we can conclude that despite the substantial costs associated with retention, its effect on immediate dropout seems negligible.

In contrast to immediate student dropout, the negative effect of retention on the probability to be observed at the Bachelor level is more pronounced. Columns (1) and (2) of Table 3.6 suggest that the effect ranges from -12 to -15 percentage points and that the effect is statistically significant. The estimates become less precise and insignificant, when larger observation windows with more flexible models are considered. The reason for the disparity between immediate dropout and later dropout is threefold. First, students might enroll into the third semester and benefit from their student status while looking for outside opportunities. Second, students might update the costs of repeating only after having started their second attempt. Third, students might fail the ASY for a second time; however, the chance of passing the second attempt is high (approximately 90%).

Overall, the analysis of the dropout effects suggests that, first, immediate dropout rates are *ceteris paribus* not significantly higher for retained students, which is, however, partly due to large standard errors. Moreover, given that retained student have to complete an additional year and face the risk of a second failure, we consider the share of retained students that is never observed at the Bachelor level moderate. These findings are suggestive of the high utility that students receive from stay-

[42] This property is a general problem of the local-linear estimator when applied to binary outcomes (Frölich, 2006).

ing enrolled despite their failed first attempt, which is supposedly due to the high earnings prospects of graduates from the University of St. Gallen.

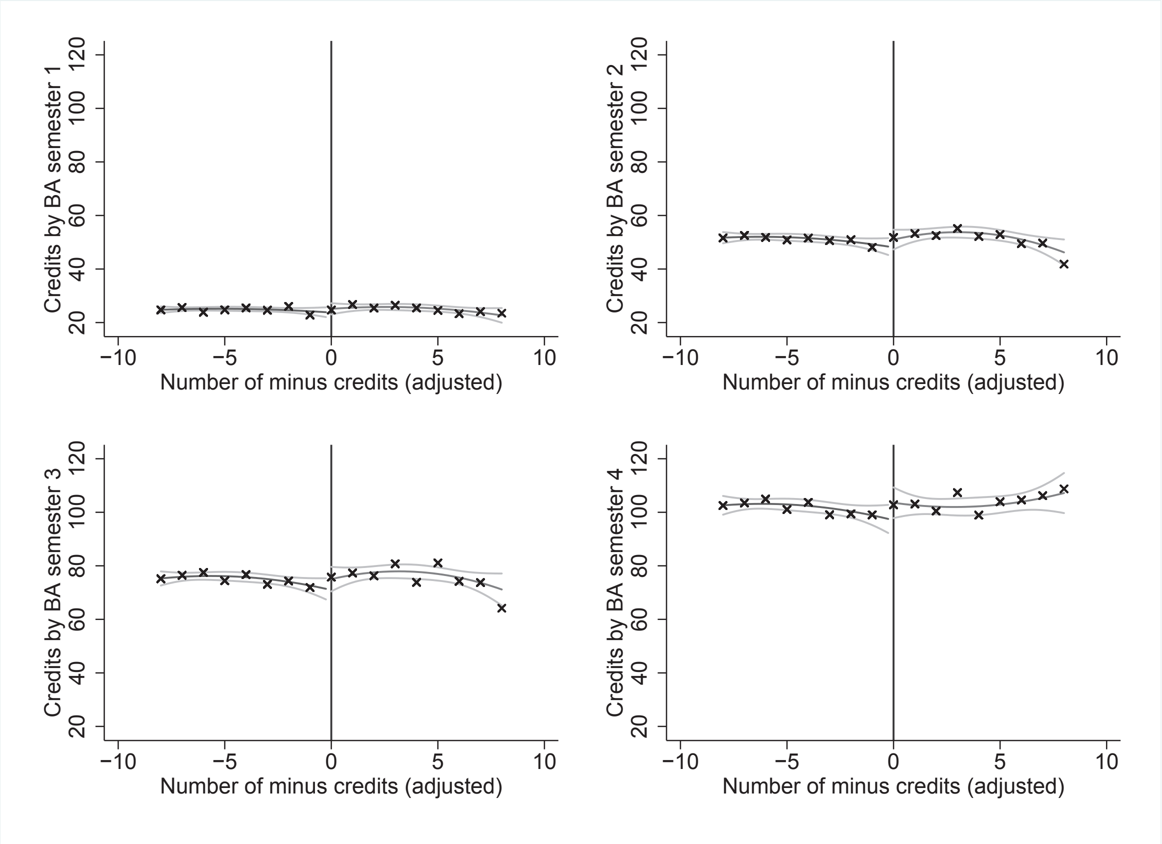
3.5.2 The effect of repeating on academic outcomes

The model estimates for the effect of repeating on academic outcomes are based on the sample of students who are observed at the Bachelor level. We start by examining major choice, which is the first decision students face after their successful completion of the ASY. Major choice is a relevant outcome for two reasons. First, major choice can determine a student's human capital formation and future earnings. Second, differences in choices of major between individuals in the treated group and individuals in the control group should be accounted for in the further analysis as students' performance (grades, credits) might differ across majors.

The lower panel of Table 3.6 investigates the effects based on various specifications of RDD models. The table shows significant differences in major choice for some specifications. Looking at columns (1), (2), and (5), it appears that retained students are on average 10–18 percentage points more likely to favor Economics as their major. The other point estimates are of similar magnitude, but statistically insignificant (again, supposedly, caused by the added model flexibility). No significant effects for other majors exist. Thus, it seems that retention has an effect on the subsequent choice of the study path. Since we neither observe students' preferences, nor changes thereof, we cannot say much about the underlying reasons for this pattern. We interpret this finding as a result of a continuous updating process throughout the ASY, during which students re-evaluate their preferences.

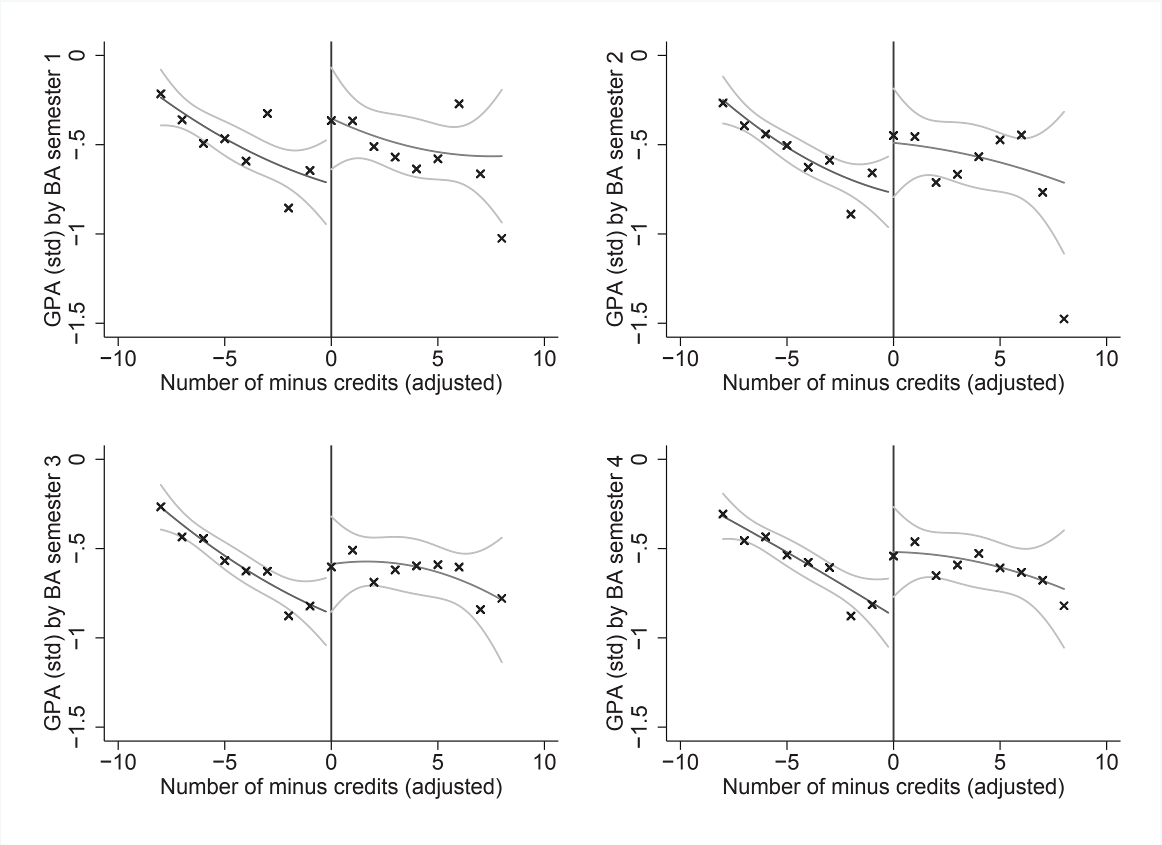
We next turn to the outcomes that measure academic performance. The patterns of accumulated credits as well as grade performance over the first four semesters at the Bachelor level are illustrated in Figures 3.4 and 3.5. Again, quadratic regression estimates for retained and non-retained students are depicted in all figures. Figure 3.4 shows that – unlike in the case for the previous outcomes – the number of accumulated credits is a flat and smooth function of the number of minus credits accumulated in the ASY. Given that we interpret this outcome as a measure of study speed, retained and promoted students seem to proceed through their major

Figure 3.4: RDD estimates: Credits accumulated by the end of each Bachelor semester



The panels above provide a graphical illustration of the RDD estimates of the number of credits accumulated for the Bachelor's degree by the end of each of the first four Bachelor semesters, respectively. The green dots represent the mean outcomes within each minus credit category. The green lines display a quadratic fit to either side of the cutoff (95% confidence intervals in grey), respectively. The sample consists of all individuals in the estimation sample within a range of 8 minus credits to either side of the cutoff (n = 819).

Figure 3.5: RDD estimates: Grade point averages (GPA) by the end of each Bachelor semester (standardized)



The panels above provide a graphical illustration of the RDD estimates of standardized grade point averages (GPAs) by the end of each of the first four Bachelor semesters, respectively. GPAs are standardized at the level of all individuals who have started their Bachelor degree in the same semester. The green dots represent the mean outcomes within each minus credit category. The green lines display a quadratic fit to either side of the cutoff (95% confidence intervals in grey), respectively. The sample consists of all individuals in the estimation sample within a range of 8 minus credits to either side of the cutoff (n = 819).

Table 3.7: RDD estimates of Bachelor outcomes

Outcome	(1)	(2)	(3)	(4)	(5)
Credits (std) after 1st BA semester	2.86**	1.50	2.88	2.64	1.47
	(1.42)	(1.42)	(1.42)	(1.42)	(1.42)
Credits (std) after 2nd BA semester	4.81**	3.16	3.92	3.48	3.58
	(2.42)	(2.54)	(3.35)	(3.11)	(2.28)
Credits (std) after 3rd BA semester	5.19	4.29	5.89	5.19	4.53*
	(3.24)	(3.27)	(4.26)	(4.15)	(2.71)
Credits (std) after 4th BA semester	4.33	7.01*	5.00	4.45	4.11
	(3.99)	(3.98)	(5.17)	(5.30)	(3.19)
GPA (std) after 1st BA semester	0.35**	0.37**	0.41	0.41*	0.46**
	(0.15)	(0.19)	(0.26)	(0.25)	(0.21)
GPA (std) after 2nd BA semester	0.29**	0.30*	0.36	0.37	0.34*
	(0.15)	(0.18)	(0.25)	(0.25)	(0.22)
GPA (std) after 3rd BA semester	0.32***	0.29*	0.29	0.31	0.31*
	(0.13)	(0.16)	(0.22)	(0.23)	(0.19)
GPA (std) after 4th BA semester	0.37***	0.34**	0.36*	0.38*	0.33*
	(0.13)	(0.16)	(0.23)	(0.23)	(0.21)
Window	[1;1]	[8;8]	[12;12]	[12;12]	[12;12]
Polynomial order	0	2	4	4	NP
Observations	82	819	1,488	1,488	1,488
Covariates	No	No	No	Yes	No

RDD estimates of dropout behavior and major choice. Columns (1)-(5) display different specifications. The parametric specifications (1)-(4) are estimated using a linear probability model. Following Imbens and Lemieux (2008) the bandwidth for the local-linear nonparametric specification (NP) is determined by cross-validation. The respective estimation window for each specification is reported as the minus credit range on each side of the threshold. Covariates include the following indicator variables: Cohort dummies, male, younger than 20 by the start of the ASY, older than 21 by the start of the ASY, non-Swiss nationality, non-German mother tongue, high school St. Gallen, entrance test participation, gap year after finishing high school. All specifications control for the choice of major studies. * Significant at 10%- level, ** Significant at 5%- level, ***Significant at 1%- level. Standard errors are in parentheses.

at a rather similar pace. This also holds when only the cut-off area is investigated. If anything, retained students close to the cut-off accumulate slightly more credit points per semester. This is also supported by the results in Table 3.7 where the effect of retention on accumulated credits is mostly insignificant, and the estimated effects become less precise as the estimated models grow more flexible. Based on these estimates, there is some weak indication that just retained students have accumulated marginally more credit points already after their first Bachelor semester, and that they extend that lead until the end of their Bachelor studies. However, the effects are rather small and do not compensate for the time lost due to repeating the ASY.

Figure 3.5 shows that the relationship between the number of accumulated minus credits during the ASY and GPAs at all stages of the Bachelor is rather negative when taking each side of the cuf-off value separately. Hence, performance in the ASY appears to have some predictive power with respect to the grades achieved later on. However, the figures also indicate that there is a structural break just at the cut-off value which shows that just retained students appear to achieve better grades in their Bachelor studies than students who just passed the ASY. The quadratic regression lines perform reasonably well in estimating the jump at the discontinuity point. The regression estimates in Table 3.7 clearly support this finding. Taking all estimates together, there is sufficiently strong evidence to state that repeating the ASY leads to significant GPA improvements in the range of 0.29 to 0.46 standard deviations. The point estimates are mostly robust, but at times somewhat too imprecise to be significant when the largest window with the fourth order polynomial is used. Nevertheless, we interpret these findings as a positive causal effect of retention on educational achievement for students who were just retained. Most importantly, the positive GPA effect persists over at least four semesters.

Although we find that both major choice and academic performance after the ASY are significantly influenced by student retention, we are not able to make any statements about how retention affected academic performance had major choice not been influenced. If retained students only performed better on average because they are more likely to study Economics, and if GPAs were higher in Economics, the estimated positive effects of retention on GPAs would be upward biased. To address such concerns, we compare mean GPAs for all Bachelor semesters between the two subgroups of Business and Economics students (not shown). After controlling for minus credits in the ASY, GPAs of Economics students are lower than GPAs of Business students throughout all Bachelor semesters. Therefore, we conclude that improvements in GPAs are not an artifact of major choice, but a direct effect of retention and repetition of the ASY.

To sum up, retention and subsequent repetition appears to have a beneficial effect on the grades of students at the Bachelor level. Moreover, this effect is persistent as it lasts throughout the entire observation period. However, there is little indication of a catch-up effect in terms of study duration. Just retained students do not accumulate more credit points per semester than promoted students. They "lose" one year and thus incur higher opportunity costs of finishing their degree than students who are not retained.

3.6 Conclusion

An ever-increasing number of incoming college students is putting existing institutions of higher education in OECD countries under pressure to provide tertiary education in larger quantities while at the same time aiming to maintain their level of quality. Where law prevents these institutions from autonomous ex-ante selection of their incoming students, assessing them in the course of a "probation year" (i.e., the first year) is a feasible alternative. In particular, students are required to meet certain academic standards by the end of their first year while non-compliance leads to retention. A growing number of institutions, especially in European countries, are nowadays applying comparable frameworks.

Using administrative data from the University of St. Gallen, this paper provides empirical evidence on the dynamics and outcomes of such a system. Analyzing six freshmen cohorts from 2001–2006, we find that roughly one fourth of students drop out already before the end of their first year. This happens at different stages of the first year, and the reasons are supposedly heterogeneous. Yet, we find descriptive evidence for potential deterrence effects that lead some weakly performing students to drop out before their actual assessment. Three-quarters of students are observed to take all the required first-year exams and form our main estimation sample (a selected sample). Accounting for the endogeneity of students' retention status in our sample by using a regression discontinuity design, we argue and show that students who perform just below the retention threshold are sufficiently comparable to students who just pass the first year. Within this selected group, we locally estimate the causal effects of being retained (and subsequently having to repeat the full year) on the subsequent dropout probability, on the choice of major studies and on subsequent educational outcomes measured up to four semesters of Bachelor studies.

We find the following results: Visual presentations confirm that retention increases immediate dropout of students after the first year. However, the regression results suffer from relatively large standard errors and are thus not significant. Beyond that, retained students are significantly less likely to ever be observed at the Bachelor level, which reflects the combined effect of immediate dropout as well as forced dropout due to failing the ASY twice. In addition, retention tends to influence the choice of major studies in favor of economics. Regardless of that choice, the effects of retention on subsequent academic performance seem favorable for the policy and persist throughout the Bachelor: by the end of the fourth Bachelor semester, retained students show on average significantly higher GPAs than their non-retained comparison group. At the same time, however, we do not find much evidence for increased study speed, i.e., catch-up effects cannot be detected. Thus, from a policy perspective, retention in higher education appears to be a reasonable measure to improve academic performance, at least when the focus of the policy is on the better performing among the retained. Yet, it comes at the cost of an additional year that students spend in education.

This study has several limitations. First, the local nature of our identification strategy limits the validity of our results to students who perform close to the minimum passing requirements as set by the university. The vast majority of students in our sample performs significantly better than required, and we cannot make any statements about the effects that retention would have on this group. Nevertheless, we argue that the subgroup that we investigate is the most relevant one from a policy perspective. Retention policies are designed to improve the academic performance of students with academic deficiencies that can presumably be straightened out. The relatively low dropout rate also confirms that most students are indeed willing to pursue a second attempt. Second, the limited number of observations does not allow us to look at effect heterogeneity, for example, across males and females or younger and older students, while such analyses would certainly provide additional insights about heterogeneity in learning behavior (Tinto, 1975). Third, we study retention effects in a particular setting. Institutions of higher education are heterogeneous in terms of the subjects that they offer, the type of students that they attract as well as their specific rules of student assessment. Overall, graduates from the University of St. Gallen have good future job and earnings prospects. Hence, we can expect students to accept higher costs (in monetary terms as well as in terms of effort) before they decide to drop out. Other institutions could attract different types of students where retention might have a stronger (or weaker) effect on motivation and academic improvement, respectively. In this light, further studies from other institutions are needed to improve our knowledge about retention effects in higher education. Further studies should also investigate the pathways through which retention affects dropout behavior and educational performance.

Appendix

3.A Figures and tables

Figure 3.A.1: Time line: Institutional setup

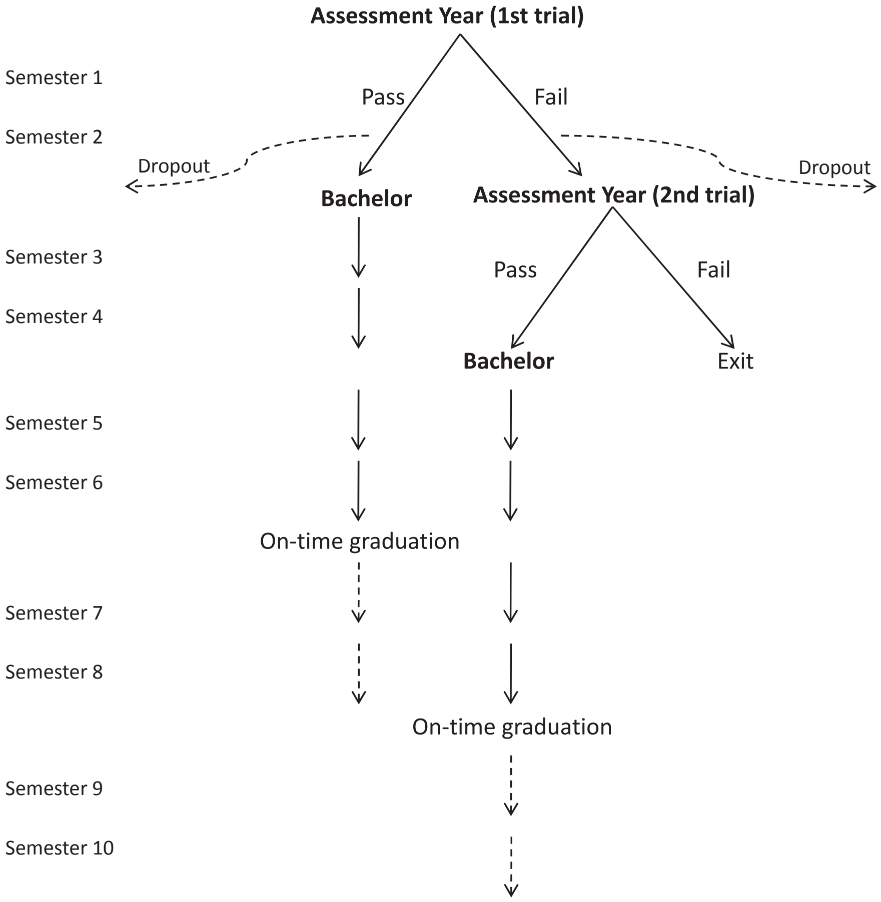


Table 3.A.1: Graduation statistics for Switzerland

Graduation	Average	2007	2008	2009	2010
Total (N=82233)	20,558	17,797	20,205	21,230	23,001
Women [%]	51.8	49.0	51.0	52.8	54.4
Foreigners [%]	16.8	16.1	17.4	17.0	16.5
Business Admin. or Econ. (N=12,258)	3,065	2,904	2,963	3,009	3382
in % of Total	15.0	16.3	14.7	14.2	14.7
Women [%]	31.2	30.4	30.3	31.2	32.8
at University of St. Gallen (N=3,640)	910	903	903	881	953
in % of Total	*4.5*	*5.1*	*4.5*	*4.1*	*4.1*
in % of Business Admin. or Econ.	*29.8*	*31.1*	*30.5*	*29.3*	*28.2*
Women [%]	19.0	18.8	18.8	19.1	19.2
Foreigners [%]	16.8	16.5	16.5	16.7	17.6

Graduation consists of Licentiate, Bachelor or Masterin Switzerland. All Percentages are rounded to one decimal place. Source: Federal Administration of Switzerland.

Table 3.A.2: Capacity constraints at the university due to high amount of entering students.

Year	No. of Students	ASY students
1990	3,908	582
...	...	...
2000	4,701	843
2001	4,938	971
2002	4,917	953
2003	4,852	900
2004	4,569	789
2005	4,508	954
2006	4,915	1022

Table 3.A.3: Typical curriculum of a student

Calendar Week	38	45-46	46		50	51	3-7						10
1st Semester	Start	register to exams	Assignments			End	A Exams						get grades
Exam			LS		CT		Econ	BA	Law	Math			
Credits			3		2		5.5	5	5.5	3.5			
Calendar Week	8	15-16	50-15	16	16-21	21	25-29						35
2nd Semester	Start	register to exams	Assignments			End	B Exams						get grades
Exam			Essay	BA*	LS		Econ	BA	Law	Math	CT	FL	
Credits			5	2	3		5.5	5	5.5	3.5	2	4	

LS	Leadership skills
CT	Critical Thinking
BA	Business Administration
BA*	Business Administration: Case study
FL	Foreign Language
Assignments	Essays or oral assignments during the semester
Exams	Written exams during the exam period

Table 3.A.4: Descriptive statistics: Entering first-year students by year

	Obs.	2001-6	2001	2002	2003	2004	2005	2006
Background Characteristics								
Male	3762	73%	75%	74%	75%	74%	72%	71%
Age $<$ 20	3762	12%	8%	9%	14%	14%	13%	17%
Age 20/21	3762	65%	64%	64%	65%	68%	66%	61%
Age $>$ 21	3762	23%	29%	26%	21%	18%	20%	22%
Foreign nationality	3762	22%	18%	24%	25%	26%	23%	20%
High school St. Gallen	3762	15%	14%	14%	17%	17%	15%	16%
Entrancetest	3762	16%	13%	17%	21%	18%	14%	14%
Types								
1st sem: Not all exams	3762	6%	13%	4%	4%	4%	5%	3%
1st sem: MC $>$ 12	3762	10%	13%	7%	13%	13%	11%	11%
1st sem: Voluntary dropout	3762	1%	2%	0%	0%	1%	0%	0%
2nd sem: Not all exams	3762	2%	2%	2%	3%	2%	3%	1%
2nd sem: Accounting failed	3762	2%	1%	4%	5%	1%	1%	1%
2nd sem: All exams	3762	79%	77%	82%	75%	80%	79%	83%
Minus Credits								
MC $>$ 0 in first year	3762	63%	69%	65%	62%	59%	63%	58%
# of MC in first semester	3762	5.12	5.04	4.71	4.57	5.24	5.70	5.41
# of MC in first year	3762	8.43	8.02	8.59	6.91	8.87	10.58	7.67
Retention								
Fail: MC $>$ 12	3762	10%	10%	13%	8%	8%	13%	8%
Fail: Credits $<$ 60	3762	3%	7%	1%	4%	2%	2%	1%
Fail: Both reasons	3762	16%	16%	12%	16%	17%	18%	15%
Fail: Total	3762	29%	32%	27%	28%	28%	33%	24%
Repetition								
Repeater	3762	17%	15%	17%	16%	17%	21%	15%
Bachelor								
Bachelor started	3762	82%	75%	83%	83%	84%	82%	86%
Bachelor $\leq$ 4 semesters	3083	36%	47%	49%	38%	31%	25%	23%
Bachelor $\leq$ 5 semesters	3083	60%	68%	71%	62%	58%	54%	47%
Bachelor $\leq$ 6 semesters	3083	83%	87%	88%	86%	84%	80%	74%
# obs		3762	794	604	569	477	646	672

The sample includes all first-year students with German mother tongue entering in 2001 - 2006 into the Business/Economics track. The last three rows (Bachelor $\leq$ 4/5/6 semesters) are only specified for students starting a Bachelor degree.

4. After-School Care and Parents' Labor Supply

Christina Felfe, Michael Lechner, and Petra Thiemann

Abstract

Does after-school care provision promote mothers' employment and balance the allocation of paid work among parents of schoolchildren? We address this question by exploiting variation in cantonal (state) regulations of after-school care provision in Switzerland. To establish exogeneity of cantonal regulations with respect to employment opportunities and preferences of the population, we restrict our analysis to confined regions along cantonal borders. Using semi-parametric instrumental variable methods, we find a positive impact of after-school care provision on mothers' full-time employment, but a negative impact on fathers' full-time employment. Thus, the supply of after-school care fosters a convergence of parental working hours.

JEL codes: J13, J22, C14
Keywords: Childcare, parents' labor supply, semi-parametric estimation methods

4.1 Introduction

Although mothers' labor market participation increased strongly during the 21st century, a substantial gender gap in work hours of mothers and fathers remains. In 2009, the average employment rate among women with children under the age of 15 amounted to 66.2% in OECD countries (OECD, 2012), but only a minority of these women worked full-time (44.6%). 26.1% of these women worked 50–90% (3-4 days per week), and 29.4% worked less than 50%. In contrast, a large majority of men with children under the age of 15 worked full-time (78.4 %). These gender differences partly arise from differential childcare responsibilities within families (OECD, 2001).

This paper provides empirical evidence on the effect of after-school care provision as a policy to promote mothers' labor supply. Many developed countries currently expand the public[43] supply of all-day schools and after-school care (Kamette, 2011), given the existing evidence on negative consequences of a reduced workload for women's career opportunities (Waldfogel, 1997, Bratti et al., 2005, Felfe, 2012). In addition to gender equality arguments, these policies follow at least two motivations for public intervention. First, individuals do not necessarily account for public returns to their labor supply. They might thus undersupply labor from a social perspective, especially when childcare costs are high.[44] Second, after-school care facilities face in general high setup-costs, which hampers market entry for private providers. On the contrary, public providers can save costs by exploiting existing infrastructure, e.g. school facilities.[45] Yet, little evidence on the impact of public supply of after-school care on parents' labor supply exists.

[43] We use the term "public" childcare interchangeably with "publicly regulated" childcare. In other words, public childcare slots do not necessarily need to be publicly financed. For details on the regulation and financing scheme of public childcare in Switzerland, our country under study, see Section 4.2.

[44] Blau and Currie (2006) provide the following examples of labor externalities, among others: Working mothers might rely less on public assistance, and they might serve as positive role models for their children. Moreover, employment of highly qualified (college-educated) parents might generate positive human capital externalities (Moretti, 2004).

[45] In addition, Blau and Currie (2006) mention information asymmetries about the quality of childcare as a rationale for public intervention.

Identifying a causal effect of childcare availability on parents' labor supply is challenging, as supply and demand for childcare are simultaneously determined at the local level in at least three ways. First, parents might influence childcare provision according to their preferences, by lobbying or voting for policies that enhance childcare supply. Second, childcare providers might choose locations with high childcare demand and high labor supply; similarly, parents with higher propensity to work might choose to locate in regions with higher childcare density. Third, municipalities might promote or subsidize childcare to attract highly qualified parents. All three mechanisms could potentially create positive correlations between childcare and labor supply, and thus upward bias a causal effect estimate.

To address the identification problem, this paper exploits legal differences in after-school care enforcement at the state (cantonal) level in Switzerland. These cantonal differences generate variation in childcare provision at the municipality level. Recent institutional changes, which address a pronounced gender gap in parents' work hours, provide the background for the analysis.[46] To promote maternal labor supply, the Swiss federal government started to subsidize the expansion of extra-familiar childcare in 2003.[47] In addition, by changing their school laws, several cantons began to enforce after-school care provision at the municipality level around the same time. Efforts at the federal, cantonal, and municipal level are complementary, as federal subsidies are contingent upon the approval and the support of the canton and the municipality. This study uses a unique database, which contains the number of after-school care slots for all 2,596 Swiss municipalities for the year 2010. We combine this information with individual-level data from the Swiss census 2010, in particular employment outcomes, demographic characteristics, and family structure. Furthermore, we enrich the dataset with municipality characteristics, provided by

[46] In 2010, in only 11% of all two-parent families with primary school children (4-12 years old) both parents worked full-time, in 47% the mother worked part-time and the father worked full-time, and in 28% the mother was not employed at all, but only the father worked full-time. These numbers are based on own calculations using the Swiss Structural Survey. Please refer to Section 4.4 for more details on the data.

[47] On February 1, 2003, the Swiss government launched a national program subsidizing new childcare facilities as well as expansions of existing childcare facilities. By 2010, this program has led to an increase of the supply in after-school care slots by about 25,000 slots.

the Swiss Federal Statistical Office, in particular demographic structure and voting outcomes.

The analysis considers small geographic areas that are homogenous in terms of employment opportunities (henceforth “local labor markets”, or “LLMs”), but that are divided by a cantonal boarder. Within these areas, we regard cantonal enforcement of childcare supply as exogenous to individuals’ labor supply decisions, conditional on a set of regional and individual control variables. Thus, cantonal enforcement serves as an instrumental variable (IV), and allows us to identify a local average treatment effect (Imbens and Angrist, 1994) of childcare availability in a family’s municipality of residence on employment indicators. Multiple papers exploit geographic borders to uncover the effects of policy interventions (Card and Krueger, 1994, Holmes, 1998, Black, 1999, Pence, 2006). These papers argue that policies change abruptly at the border, but the economic environment changes only little. This paper departs from the two-stage-least-squares methods typically applied in these contexts, and instead uses the semi-parametric methodology by Frölich and Lechner (2010), which relies on fewer functional form assumptions. In particular, the approach incorporates control variables in a non-linear way, and allows for heterogeneous treatment effects across the LLMs. Accounting for regional heterogeneity is important in our context, given that responses to childcare provision depend on differences in initial childcare levels, and on differences in the institutional environments (Fitzpatrick, 2012). To aggregate estimates across LLMs, we propose specific aggregating schemes.

This paper contributes to a broad literature that analyzes the consequences of childcare provision for mothers’ labor supply. Several studies focus on the impact of childcare provision for children in preschool age. In his seminal paper, Gelbach (2002)[48] uses quarter of birth indicators as instruments for childcare attendance in the US. He finds that providing public childcare free of charge stimulates employ-

[48] Several earlier papers study the impact of childcare prices on mothers’ labor supply. Most of these papers estimate structural parameters of utility functions to derive mothers’ labor supply elasticities and to predict the consequences of childcare subsidies (Blau and Robins, 1988, Connelly, 1992, Michalopoulos et al., 1992, Kimmel, 1998). The resulting estimates of mothers’ labor supply elasticities with respect to childcare prices vary between 0 and -1.6 for married mothers.

ment by 6–15% among married mothers and by 6–24% among single mothers. A first strand of literature supports these findings. Most identification strategies rely on regional and time variation in childcare supply (Berlinski and Galiani, 2007, Nollenberger and Rodríguez-Planas, 2011, Schlosser, 2011), or on the introduction of a price subsidy for public care (Baker et al., 2008, Lefebvre and Merrigan, 2008). A second strand of literature finds, however, that maternal labor supply is on average rather inelastic to exogenous increases in childcare availability. Only subgroups of mothers, such as single mothers or mothers in disadvantaged areas, react positively to an increase in public childcare (Cascio, 2009, Fitzpatrick, 2010, Goux and Maurin, 2010, Havnes and Mogstad, 2011). Fitzpatrick (2012) discusses the reasons for the lack of consensus in the empirical findings, using the US as example. On the one hand, the studies differ in their empirical methodologies. On the other hand, childcare policies vary with respect to their institutional and socio-economic contexts: Over the years, the subset of mothers whose labor supply reacts to an expansion of public childcare has potentially shrunk because of increasing maternal employment, delayed childbearing ages, and rising educational attainment.

To our knowledge, only one study focuses on the effects of childcare for schoolchildren (Lundin et al., 2008). The authors evaluate the effects of a price reduction of childcare for children age 0-9 years old in Sweden at a time when overall childcare coverage was already high (80%). Their results reveal positive effects of subsidized childcare on overall maternal employment. Yet, for mothers of children in the age of 5 and older, these effects are negligible.

Our analysis reveals a positive impact of after-school care provision on mothers' full-time employment, but a negative impact on fathers' full-time employment. In particular, an increase in after-school care provision by on average 8 percentage points (henceforth “ppt") leads to an average increase in mothers' full-time employment by 8 ppt. In contrast, the same increase in after-school care provision crowds out fathers' full-time employment by 10 ppt.

This study contributes to the literature in four ways. First, this paper evaluates the impact of expanding public care provision for schoolchildren in a context of low initial levels; in Switzerland, the coverage rate is on average about 9% among children in the age of 4-12. Thus, if levels have an impact on the magnitudes of the effects, our results might differ from those of Lundin et al. (2008). Second, in contrast to existing studies, this paper focuses on fathers' employment as well. Thus, the analysis sheds some light on whether extra-familiar care improves the allocation of paid work among men and women. Third, the analysis also considers the intensive margin. In the light of high maternal employment rates, but prevailing gender wage differences, this focus helps to reveal changes in labor supply at margins that are relevant for women's career opportunities. Finally, the semi-parametric IV methodology used in this study allows for effect heterogeneity and for a flexible way of controlling for observables. This is the first paper to apply such a methodology to the analysis of childcare effects.

The paper proceeds as follows: Section 4.2 provides an overview of the childcare system in Switzerland and the cantonal regulations of after-school care provision. Section 4.3 explains the identification strategy and estimation method. Section 4.4 describes the data, and Section 4.5 shows the results and a series of robustness checks. Section 4.6 concludes.

4.2 Institutional background: After-school care in Switzerland

In Switzerland, labor market outcomes of parents with schoolchildren (age 4-12) strongly differ by gender. In only 11% of families with schoolchildren, both parents work full-time, in 47% of these families, the mother works part-time and the father works full-time, and in 28% of these families, the mother does not work, and the father works full-time.

To promote mothers' labor market participation, the Swiss government has launched a federal program in 2003. This program subsidizes new childcare facilities as well as expansions of existing childcare facilities during the first three years after their establishment or expansion. Both public and private providers are eligible for the subsidy.[49] By February 2010, the program has financed 25,000 new childcare slots, which corresponds to an increase in childcare coverage by approximately 50%, according to the Federal Social Insurance Office of Switzerland (FSIO, 2010). About half of this increase is due to increases in after-school care coverage. In 2010, average coverage rates of extra-familiar care amounted to 15% among pre-school children (age 0-3) and to 9% among schoolchildren (age 4-12).[50]

Childcare coverage rates vary substantially at the cantonal level, with cantonal coverage rates ranging from 1% to 23% for pre-school children and from 1% to 43% for schoolchildren (see Figure 4.2). Where do the differences between cantons come from? Cantons differ in their policies to support childcare provision (see Table 4.A.1 in the appendix for an overview). For instance, 19 out of 26 cantons explicitly mention extra-familiar childcare as one policy to support families in their legislation; 17 cantons provide information and counseling to childcare facilities that want to apply for federal subsidies; and 15 cantons contribute financially to the provision of childcare.[51]

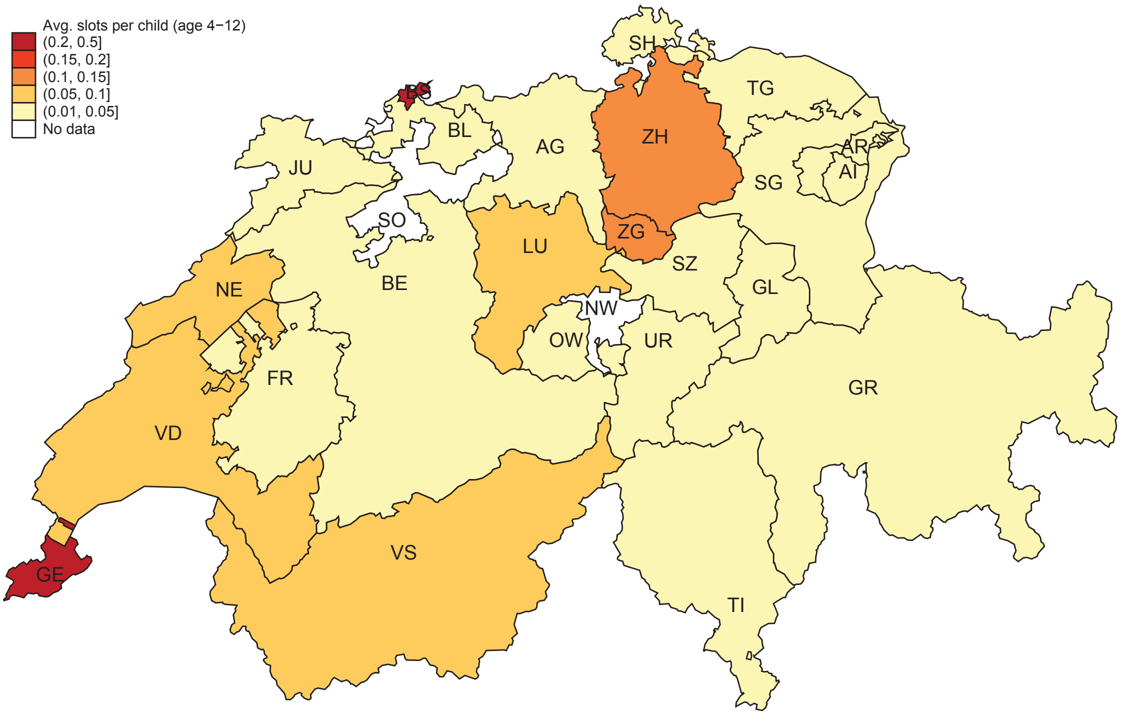
[49] The program has been launched on February 1, 2003. It is called "Federal Law on Financial Support for Extra-Familiar Childcare" ("Bundesgesetz über Finanzhilfen für familienergänzende Kinderbetreuung") and is administered by the Ministry of Social Affairs (Bundesamt für Sozialversicherung). Article 1 of the law states the purpose of the program: "The Swiss federation provides [...] childcare subsidies [...] so that parents can better reconcile family life with work and/or education" (own translation).

[50] This data stems from a recent data collection by Infras, Zurich, and the Swiss Institute of Empirical Economic Studies at the University of St. Gallen. It facilitates for the first time a national overview of childcare availability in Switzerland and thus allows for transparency and comparison across and within cantons for the year 2010. For details, please refer to Felfe et al. (2013). Unfortunately, no data for previous years are available, which prevent us from any identification based on the expansion of after-school care supply over the years since implementation of the program.

[51] Childcare costs are generally borne by parents. Public demand subsidies (by the canton or the municipality) are available to low-income families and are paid independently from the childcare provider (in other words, public subsidies can be used to pay a slot in a publicly or a privately organized childcare institution). The availability and amount of these public subsidies, however, varies greatly across and within cantons. Unfortunately, so far no reliable data on the availability or the amount of public demand subsidies exist. Therefore, our study can only provide estimates for the impact of the availability of childcare slots. Yet, comparing Tables 4.A.1 and 4.A.2 reveals that financial support by the canton is not systematically correlated with cantonal regulations regarding the supply of after-school care (our instrumental variable, described in more detail in Section 4.3).

Coverage rates vary not only across cantons, but also within cantons. For instance, in the canton Zurich, 1% of schoolchildren live in a municipality without after-school care coverage, while 54% of schoolchildren live in a municipality with a coverage rate of more 10% (see Figure 4.2). In the canton of Bern, these shares correspond to 47% and 2%, respectively. Heterogeneity within cantons comes from municipalities' discretion in the provision of childcare, but this discretion in turn depends on cantonal laws. Depending on the legal setup, either the canton alone, either the municipality, or both regulate, license, and supervise the provision of childcare facilities (see Table 4.A.1).

Figure 4.1: Coverage rates of after-school care by cantons 2010

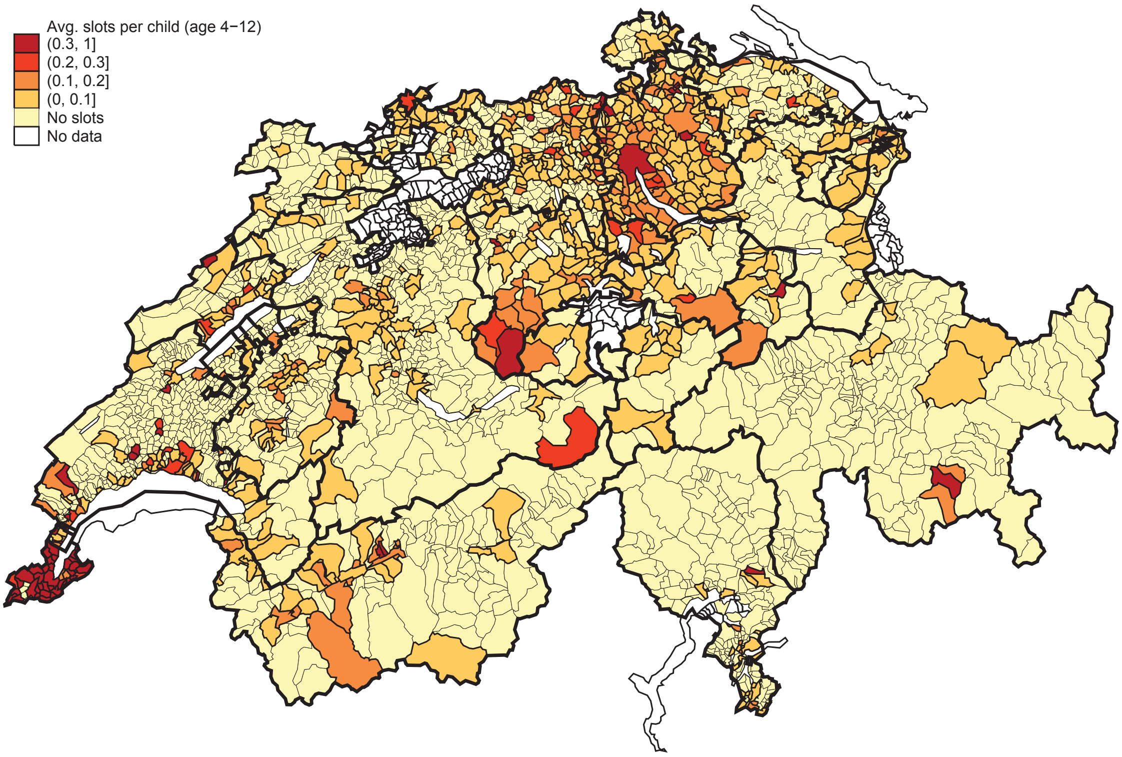


Source: Own calculations based on the population survey 2010 and childcare database.

In the course of reforming their cantonal legislation, in particular their school laws,[52] several cantons have enforced the provision of supplementary care for schoolchil-

[52] On May 21, 2006, the Swiss population and the Council of States accepted the revision of the education article in the Swiss constitution. Consequently, all cantons are obliged to regulate certain

Figure 4.2: Coverage rates of after-school care by municipalities 2010



Source: Own calculations based on the population survey 2010 and childcare database.

dren during lunchtime and during the afternoon. These reforms address the gap between supply of and demand for after-school care. What motivates such a public intervention? Public institutions seem more effective in providing after-school care than individuals or private providers, for at least two reasons. First, a private solution at the family level comes at much higher costs per child than a solution in an after-school care facility where staff-child ratios are comparable to teacher-child ratios in primary schools (1:15 and higher). Second, after-school care facilities face rather high setup costs (e.g., provision of the infrastructure). Because of these high initial investment costs and because of uncertain returns, market entry can be difficult for private providers. On the contrary, publicly supported providers can more easily access existing public infrastructure (e.g., school infrastructure), and can more easily pool their risks (e.g., through public funding of multiple institutions).

Table 4.A.2 in the appendix provides an overview of the cantonal school laws.[53] Geneva was the first canton to enforce after-school care provision. Since 2007, cantons in the German-speaking region are slowly catching up. By 2010, the year of our empirical analysis, Bern, Solothurn and Zurich have established the enforcement of supplementary after-school care. By then, also Aargau, Basel Country, and St. Gallen have incorporated lunch provision as one goal in their cantonal legislation, but not, however, care during afternoon hours. Further cantons such as Basel City, Graubünden, Lucerne, Neuchâtel, and Schaffhausen have included the enforcement of after-school care provision in their school laws only after 2010.[54]

elements of the education system (e.g. school entrance age, length of mandatory schooling). In addition, on August 1, 2009, the "HarmoS-Konkordat" came into force, which aims at harmonizing the Swiss educational system.

[53] Table 4.A.2 is based on careful reading and interpretation of the cantonal laws. We explicitly distinguish between laws only referring to childcare provision as one policy to promote families, requiring an inquiry of supply and demand of childcare, or enforcing the supply of sufficient childcare facilities. Only the latter is interpreted as legal enforcement.

[54] Basel Country and Graubünden did so in 2011, Lucerne and Schaffhausen in 2012, and Neuchâtel plans to do so in 2015.

4.3 Econometric framework

4.3.1 Identification

To account for potential endogeneity of childcare supply with respect to parents' labor market outcomes, we implement an instrumental variables (IV) strategy; that is, we exploit regional variation in childcare availability that arises from differences in cantonal enforcement of childcare supply. In the Swiss setting, we assume exogeneity of cantonal enforcement to parents' labor supply decisions under specific conditions (discussed below). The treatment is high childcare coverage in an individuals' municipality of residence, and the control condition is low childcare coverage (see Section 4.3.2 on the categorization into high versus low treatment levels). The corresponding parameter of interest is the local average treatment effect (LATE), which is the effect of high childcare coverage on individuals living in "complier municipalities" (Imbens and Angrist, 1994). Complier municipalities are those municipalities whose coverage is high if and only if their canton enforces childcare supply. As schoolchildren can only attend school without charge in their canton of residence, and after-school care takes place mostly in schools, our treatment definition relies on a family's residence (and not, for example, on the parents' workplace).[55]

The LATE has a causal interpretation only when two conditions hold. First, cantonal legislation must causally influence childcare supply at the municipality level, and second, cantons must influence labor market outcomes only through the channel of childcare legislation and not through alternative channels ("exclusion restriction").

We argue that the first assumption holds because of the institutional setting in Switzerland. Some cantons legally enforce the provision of after-school care; while others at most mandate an inquiry of the supply and demand of after-school care (see Section 4.2). As a result, municipalities exert differential effort to ensure the provision of sufficient after-school care slots, depending on the canton they belong. Furthermore, as federal funds for after-school care projects are contingent upon the

[55] Unfortunately, our data do not contain information on after-school care take-up.

approval and support of the canton, cantons exert direct influence on the number of institutions that extend their childcare supply. A cross-cantonal comparison of coverage rates provides evidence for the impact of cantonal legislation: In most of the LLMs considered here (see below for a detailed description), the coverage rates in cantons with enforcement lie on average above the coverage rates in cantons without enforcement (see 4.1, column 6, for unconditional means, and Section 4.5.1 for conditional means).

The second assumption, that is, the exclusion restriction, is more difficult to justify and unlikely to hold in general, owing to two concerns: First, cantons differ in their industry structure and thus in their employment opportunities. Second, cantonal laws reflect the preferences of the local population. Thus, cantonal enforcement might occur in regions with better employment opportunities, or in regions with stronger preferences for policies that enable parents to work. Both of these associations most likely bias the effect of after-school care upwards. To address the two above-mentioned concerns, we therefore follow Frölich and Lechner (2010) and restrict the analysis to confined regions along cantonal borders, in particular to economically integrated local labor markets. Appendix 4.B describes the construction of LLMs in detail and provides a map of the LLMs (see Figure 4.B).

More precisely, to address the first concern – individuals residing on different sides of a cantonal border have different employment opportunities –, we define an LLM as "integrated" if all individuals residing in an LLM have approximately the same job opportunities. Thus, for any two individuals residing in the same LLM, the cost of commuting to each potential workplace must be approximately the same. We ensure this condition by setting the maximum difference in commuting times between any two individuals and to any potential workplace to half an hour; therefore, for any two individuals living in the same LLM, the choice of workplace should not depend on their canton of residence.[56]

[56] Section 4.5.2 discusses furthermore the assumption that individuals within the LLM face equal employment opportunities by providing an overview of commuting time to major economic hubs from the municipalities on both sides of the cantonal border within the LLMs (see Table 4.A.8 in the Appendix).

Table 4.1: Local labor markets: Size, political preferences, and after-school care

LLM	Canton	IN/OUT*	# of munici-palities	Population share of canton	% votes in favor of referendum**	After-school care (slots per child)
	(1)	(2)	(3)	(4)	(5)	(6)
1	BE (IV=1)	IN	50	8.00%	43.50%	0.012
		OUT	336	92.00%	55.10%	0.027
	LU (IV=0)	IN	53	37.40%	38.70%	0.027
		OUT	34	62.60%	46.30%	0.082
2	ZH (IV=1)	IN	13	3.40%	53.40%	0.09
		OUT	158	96.60%	53.40%	0.145
	LU (IV=0)	IN	14	43.20%	51.50%	0.09
		OUT	73	56.80%	37.30%	0.04
3	ZH (IV=1)	IN	24	9.30%	49.50%	0.087
		OUT	147	90.70%	53.80%	0.149
	AG (IV=0)	IN	60	35.60%	47.00%	0.059
		OUT	160	64.40%	42.40%	0.044
4	ZH (IV=1)	IN	60	26.50%	47.20%	0.09
		OUT	111	73.50%	55.70%	0.164
	AG (IV=0)	IN	40	27.30%	49.30%	0.07
		OUT	180	72.70%	42.00%	0.042
5	ZH (IV=1)	IN	79	22.60%	49.60%	0.107
		OUT	92	77.40%	54.50%	0.155
	SH (IV=0)	IN	25	99.40%	47.00%	0.025
		OUT	2	0.70%	27.40%	0
6	ZH (IV=1)	IN	73	14.80%	45.80%	0.076
		OUT	98	85.20%	54.70%	0.157
	TG (IV=0)	IN	28	35.80%	40.10%	0.031
		OUT	52	64.20%	39.10%	0.027
7	ZH (IV=1)	IN	22	5.50%	48.20%	0.068
		OUT	149	94.50%	53.70%	0.148
	TG (IV=0)	IN	49	58.80%	41.20%	0.035
		OUT	31	41.20%	37.00%	0.02
8	ZH (IV=1)	IN	22	5.50%	48.20%	0.068
		OUT	149	94.50%	53.70%	0.148
	SG (IV=0)	IN	10	13.70%	41.00%	0.019
		OUT	75	86.30%	41.60%	0.017

*IN/OUT refer to the municipalities within a canton that do/do not belong to the respective LLM. Abbreviations of cantons: AG: Aargau, BE: Bern, GR: Graubünden, TG: Thurgau, LU: Luzern, SG: St. Gallen, SH: Schaffhausen, ZH: Zurich; **Share of votes in favor of the referendum on maternity benefits on 26/09/2004.

Furthermore, to address the second concern – individuals residing on different sides of a cantonal boarder have different preferences for cantonal enforcement of childcare supply – we ensure that contrasts in political choices at the cantonal level do not result from contrasts in political preferences within the LLMs. We therefore impose three conditions. First, the population inside an LLM must not comprise the majority of any of the cantonal populations (see Table 4.1, column 4, for evidence).[57] Otherwise, an LLM's population could determine cantonal laws. Second, inside any LLM, the populations on both sides of the cantonal border should have similar preferences related to work and family. Results of a recent referendum on maternity benefits at the municipality level provide suggestive evidence for this condition (see Table 4.1, column 5). Indeed, voting results on this referendum are rather similar on both sides of the cantonal borders within each LLM. Third, regions outside the LLM should drive cantonal differences in legislation. Again, the referendum on maternity benefits provides evidence on this condition. As Table 1 suggests, on at least one side of the cantonal border, the remaining cantonal population outvotes the population living in any municipality within the LLM (see Table 4.1, column 5).

4.3.2 Estimation

The estimation proceeds in two steps. First, we estimate the LATE for each LLM separately ("within-LLM IV"). Second, we aggregate the effect over all LLMs. The first step accounts for effect heterogeneity across local labor markets. The second step increases the precision of the estimates. Effect heterogeneity is an important concern in this application, as the true effect of after-school care provision on parental labor supply may vary across individuals and LLMs. On the one hand, individuals' reaction to a change in available after-school care depends both on observable characteristics (e.g., education or income), and on unobservable characteristics (e.g., attitude towards sending their child to formal care). On the other hand, the treat-

[57] There are two exceptions where the cantonal area included in the LLM covers more than 50% of overall canton (see LLMs 5 and 7). In these cases, however, the wedge in the cantonal regulations regarding after-school care provision is caused by the other cantons (in both cases the municipalities outside the area included in the LLM in the canton Zurich).

ment effect may vary depending on the institutional context. For instance, depending on the level of after-school care supply, different types of individuals might decide to use after-school care. Since the level of after-school care supply varies strongly across LLMs (see Table 4.1, column 6), treatment effects are most likely heterogeneous in our application.

The within-LLM IV estimator combines the estimation approach by Frölich (2007), which extends the LATE framework by Imbens and Angrist (1994) to allow for control variables by matching on the propensity score, with the findings of a large-scale simulation study by Huber et al. (2013). The estimator corresponds to a ratio of two matching estimators; that is, the effect of the instrument on the outcome is divided by the effect of the instrument on the treatment.[58] Since this method relies on a binary treatment, we define a cut-off that categorizes municipalities in areas with relatively high after-school care coverage – treated municipalities – and areas with relatively low after-school care coverage – control municipalities. Given the high variation in after-school care coverage between LLMs (see Table 4.1, column 6), a single cut-off for all LLMs would result in a rather unequal distribution of treated and control areas within LLMs. We therefore define separate cut-offs for each LLM. The LLM-specific median as cut-off guarantees a similar number of treated and control observations in each LLM. The resulting cut-off coverage rates vary between 0.4% and 8.1% (see Table 4.A.3 in the appendix). The difference between the average care coverage in municipalities below and equal to the cut-off and the average care coverage in municipalities above the cut-off – the treatment intensity – amounts to 8 ppt on average, but varies across LLMs (between 5 and 11 ppt, see Table 4.A.3 in the appendix).

After estimating the effects for each LLM separately, we aggregate the different effects to increase precision. Since the IV estimates are the effects for "compliers", that

[58] To compute the two matching estimators we use the bias-adjusted-radius-propensity-score matching approach. This estimator uses a parametric propensity score to remove the effect of observable confounders that might jeopardize the validity of the instrument. By using a parametric (probit) model for the link between instruments and instrument confounders only, and being otherwise fully nonparametric, such estimators avoid the 'curse of dimensionality' which is inherent to all non-parametric procedures, but at the same time retain most of their flexibility. The results on the probit estimations for each LLM are shown in Table 4.A.11 in the appendix.

is, the effects for individuals living in “complier municipalities” (see Section 4.3.2), our preferred weighting scheme is based on the number of compliers in the respective LLM.[59] In addition, we propose three alternative weighting schemes, based on the following populations: first, based on the number of compliers, but using only those LLMs where the estimates are within the logical range (in other words, where the effect of cantonal enforcement on childcare coverage is positive); second, based on the number of observations of the respective LLM; and third, based the number of observations, but using only those LLMs for which the estimates are within the logical range. Inference is based on bootstrapping and the quantile method, that is, bootstrapping the effects and considering their distribution to obtain significance levels. We implement the bootstrap as a block bootstrap taking into account the possible correlation of individuals within the same municipality.

4.4 Data

The analysis requires information on childcare coverage, on parents’ labor supply, and on individual and regional control variables. Data on childcare coverage comes from a newly established database that contains information on the exact number of childcare slots for children in school age at the municipality level for the year 2010. Individual-level data stems from the Swiss structural survey 2010 (’Strukturerhebung 2010’). This survey supplements the Swiss census 2010 and contains information on employment status, work hours, and socio-demographic characteristics for around 200,000 randomly selected individuals among all permanent residents age 15 years and older. Information on the municipality of residence allows us to merge information on the availability of after-school care at the municipality level. We add information on the demographic and socio-economic composition of the municipality from 2010 and on the local results on the referendum on maternity benefits from 2004. The Swiss Federal Statistical Office provides all of these variables. We restrict our sample to the area covered by the LLMs, and then further to all working

[59] Estimated by the denominator of the IV estimator times the number of observations.

age (18-62 years old) men and women who have at least one child in the age between 0 and 12.[60] The samples correspond to 10,133 men and 10,875 women.

Our outcome variable is parents' labor supply. We distinguish between the extensive margin – whether parents work at all – and the intensive margin – whether parents work full-time (more than 36 hours per week) or part-time (less or equal than 36 hours per week). We also distinguish between less than 20 hours per week (low part-time), between 21 and 27 hours per week (intermediate part-time), and between 28 and 36 hours per week (high part-time).

Tables 4.A.4 and 4.A.5 provide descriptive statistics on the labor supply for the female and male samples. 70% of all women in our sample are employed. Only 10% of these women work full-time. The majority work on a low part-time basis (38%), followed by an intermediate part-time basis (16%). The majority of men, by contrast, work full-time (89%), and only few men work part-time (8%). In line with the expectation that a higher coverage rate of after-school care stimulates mothers' labor supply, mothers residing in treated areas are on average more likely to work (72% versus 68%). Furthermore, they are more likely to work full-time (11% versus 9%), and more likely to work part-time (61% versus 58%). In contrast, men residing in treated areas are slightly more inclined to work part-time (9% versus 7%), but slightly less inclined to work full-time (88% versus 91%).

Do treated and control municipalities differ along further dimensions? As Tables 4.A.4 and 4.A.5 in the appendix display, men and women living in treated areas are slightly better educated, but have fewer children on average. In addition, women living in treated areas are slightly more likely to be divorced. Treated areas are rather urban and thus more densely populated than control areas, have a higher share of foreigners, a lower share of homeowners, and a lower share of commuters.

[60] The reason for considering men and women with children age 0–12 years old instead of men and women with children age 4–12 years old is that availability of care facilities for school-age children might influence parents' labor supply decisions already during their children's preschool age. The Swiss Structural Survey only provides us with information on children living in the household. Hence, our sample does not include parents whose children are living outside the household. Yet, given the age range of the children under study, this issue might not be troublesome. Moreover, our interest lies in the effect of after-school care provision for parents who actually need to arrange childcare.

As expected, a higher share of the population votes in favor of the referendum on maternity benefits, although the referendum did not receive a majority in these areas. These differences between the treated and control areas highlight the concern that after-school care supply is endogenous to the type of authorities and population living in a municipality, even if we restricted the analysis to the LLMs. Therefore, we account for endogeneity by using an IV approach, and in addition, we control for a set of individual and municipality characteristics (in particular, education, age, household composition, and political preferences at the municipality level, see Table 4.A.11 in the appendix).

4.5 Results

4.5.1 Main results

Table 4.2 displays the main results. Panel A and B show the effect estimates for females and males. The estimates are weighted averages of the LLM-specific estimates, where the weights correspond to the number of compliers.[61] Columns 1 and 2 display the estimates of the mean potential outcome for men and women living in complier municipalities, separately for municipalities with and without enforcement of after-school care provision.[62] Column 3 shows the estimated effect (computed as the difference between column 2 and column 1), and column 4 displays the 95% confidence interval.

Cantonal enforcement of after-school care provision induces a significant increase in after-school care availability. On average, cantonal enforcement shifts the treatment status for 46% of women (43% of men) in our sample. That is, for 46% of women (43% of men), after school care supply in their municipality of residence rises from below the LLM-specific median to above the LLM-specific median. How can we interpret the treatment in terms of coverage rates? As Table 4.A.3 shows, low

[61] The appendix (Tables 4.A.14, 4.A.15, and 4.A.16), contains results using alternative aggregation schemes. The results barely change, compared to the results in Table 4.2.

[62] For a derivation of the estimators for these potential outcomes, see Frölich and Lechner (2010).

supply municipalities – with an after-school care coverage below the LLM-specific median – offer on average three slots per 100 children. In contrast, high supply municipalities – with an after-school care coverage above the LLM-specific median – offer on average 11 slots per 100 children. Thus, after a cantonal enforcement of after-school care availability, coverage increases by on average eight slots per 100 children.

What are the consequences of such an increase in after-school care for parents' labor force participation? Overall, no statistically significant change in employment status exists, both for men and for women. Yet, a statistically significant adjustment in full-time employment for both women and men appears. An increase in after-school care by on average eight slots per 100 children leads to an increase in women's full-time employment of similar magnitude: full-time employment among the women in the sample rises from 4% to 12% on average. In contrast, full-time employment among the men in our sample decreases from 96% to 87%.

Unfortunately, the imprecision of the estimates for employment and part-time employment precludes strong conclusions on the sources of the adjustment in full-time employment. In the case of women, an increase in overall employment (by 7 ppt) and a slight decrease in part-time employment (by 1 ppt) parallel the observed increase in full-time employment. In the case of men, an increase in part-time employment (by 7 ppt, significant at the 15% significance level) and a slight decrease in employment (by 2 ppt) accompany the observed decrease in full-time employment. Yet, these estimates display the reaction of the (complier) population on average and do not allow for conclusions on individuals' switching behavior as a reaction to the treatment (i.e., whether individuals change from no employment to part-time employment, from part-time to full-time employment, or even from no employment to full-time employment).

Table 4.2: Results of IV estimations, parents with children, age 0-12 years old

	Potential outcome (weighted avg.) in complier municipalities		Treatment effect (weighted avg.)	95% CI	
	w/ cantonal enforcement	w/o cantonal enforcement			
	(1)	(2)	(3)	(4)	
Panel A) Swiss women age 18-62 with children (age 0-12)					
First stage estimates:					
Effect of instrument on treatment	0.67	0.22	0.45***	0.31	0.57
LATE estimates:					
Employment	0.77	0.7	0.07	-0.05	0.2
Full-time	0.12	0.04	0.08**	0	0.18
Part-time	0.64	0.66	-0.01	-0.14	0.11
Low part-time	0.44	0.43	0.01	-0.14	0.14
Intermediate part-time	0.14	0.16	-0.02	-0.18	0.1
High part-time	0.06	0.06	0	-0.06	0.08
Panel B) Swiss Men age 18-62 with children (age 0-12)					
First Stage estimates:					
Effect of instrument on treatment	0.66	0.24	0.42***	0.3	0.55
LATE estimates:					
Employment	0.94	0.96	-0.02	-0.1	0.02
Full-time	0.87	0.96	-0.10**	-0.21	0
Part-time	0.07	0	0.07	-0.01	0.17
Low part-time	0.03	0	0.02	-0.01	0.08
Intermediate part-time	0.01	-0.02	0.02	0	0.06
High part-time	0.04	0.01	0.03	-0.03	0.08

*significant at the 1% ; **significant at the 5%; *significant at the 10%-level. Above estimates are weighted averages of the IV estimates for each LLM (LATE), based on 10,875 observations for women and 10,133 observations for men. The weights correspond to the number of compliers in the respective LLM. The instrument is based on the enforcement of after-school care supply in the cantonal law.

4.5.2 Internal validity

This section discusses the internal validity of the implemented IV method, in particular the exclusion restriction. Apart from childcare enforcement by the canton, work incentives and employment opportunities must be independent of an individuals' canton of residence – within each LLM and conditional on a set of observable characteristics. Further observed and unobserved differences in individual and cantonal characteristics might threaten the validity of the results. Therefore, this section provides three types of checks. First, to investigate institutional differences between cantons, we compare income taxes and education systems (Tables 4.A.6 and 4.A.7 in the appendix). Second, to assess individual location selection within LLMs, we check for selective migration to areas with higher childcare density as well as differences in distances to economic hubs (see Tables 4.A.8 and 4.A.10 in the appendix). Third, to evaluate selection on observable and unobservable characteristics, we conduct a placebo test; that is, we investigate whether an increase in after-school care slots stimulates the employment of a group on which it actually should have no impact: men and women under the age of 42 without children (see Table 4.A.9 in the appendix).[63]

In Switzerland, income tax schemes fall into the jurisdiction of cantons. If income taxes are systematically lower in cantons enforcing the supply of after-school care, our estimation results might be upward biased, as incentives to engage in the labor market might be due to lower taxes and not due to higher supply of after-school care. Table 4.A.6 displays average income taxes for married couples with two children and with an annual income of 100,000 CHF (as one example) on both sides of the cantonal border inside of each LLM. In 6 out of 8 LLMs, income taxes are slightly lower in the municipalities that belong to a canton that enforces childcare provision. Yet, the differences are economically negligible (at most 1 ppt.) and thus are not likely to threaten the estimates.

[63] We restrict the placebo sample to men and women under the age of 42, as older individuals are more likely to have children who no longer live in the household. As a result, available childcare services might have affected their labor force engagement in the past. Unfortunately, our data set does not allow us to identify individuals who have children that already moved out.

Do citizens on one side of the cantonal border live systematically closer to economic hubs, which offer more employment opportunities? As we can see in Table 4.A.8, individuals residing in municipalities of the canton with after-school care enforcement live indeed closer to major economic hubs such as Zurich or Berne. On average, these individuals need to commute a quarter of an hour less to these hubs. On the contrary, individuals on the other side of the cantonal border need to commute substantially less to further important economic hubs, here represented by the capital of the second canton in the LLMs. Thus, job opportunities should be comparable for all individuals residing in the same LLM.

Selective migration into cantons with higher childcare availability constitutes one further threat to our identification strategy. If parents expect easier access to after-school care in a neighboring canton, they might decide to move. Yet, no strong pattern of migration towards areas with higher childcare supply exists (see Table 4.A.10 in the appendix). Given the relatively high costs of changing residence, compared to the uncertain benefits from slight increases in after-school care coverage, this behavior seems intuitive.

Placebo estimation results support the validity of the analysis. We estimate the effect of childcare availability on a group for which we expect no effect, that is, childless individuals. If work incentives were different to both sides of the cantonal boarder inside an LLM, employment effects for these individuals would occur. Table 4.A.9 in the appendix displays the results for Swiss men and women, age 18–42, without children. The results on all of the employment indicators are not only statistically insignificant, but also economically negligible. Thus, no relevant differences in employment incentives between cantons within the area of integrated LLMs seem to exist.

4.5.3 External validity

In addition to internal validity, we ask to which extent our findings are representative for the German-speaking area in Switzerland (external validity). Notice that the IV method yields results only for individuals in "complier municipalities" inside LLMs. We therefore first assess the similarity between the the population in complier municipalities and the overall population of the LLMs. Second, we compare the population inside the LLMs with the overall population residing in the German-speaking area of Switzerland.[64]

Regarding the first comparison (within LLMs), observable characteristics of the population in complier municipalities are not statistically different from observable characteristics of the overall population in LLMs. The latter constitutes 30% of the German-speaking population in Switzerland. For instance, both populations are similar in terms of their labor force attachment, and in terms of socio-demographic characteristics (such as age, marital status, or education). Also, in terms of expressed preferences, measured by the results of the referendum on maternity benefits, both populations are comparable. Less similarity exists between LLMs and the overall German-speaking region. As described in Section 4.3, the LLMs do not comprise the majority of the cantonal population, and hence, LLMs do not include major cities. Our sample therefore underrepresents urban areas, but represents the agglomeration and rural areas of the German-speaking area of Switzerland well. Nevertheless, the differences in socio-economic and demographic characteristics between LLMs and the overall German-speaking area are negligible.

4.6 Conclusion

This paper addresses the question whether after-school care provision can affect parental labor supply. Relying on cantonal regulations in after-school care provision as instruments, and using semi-parametric instrumental variable methods, we find that after-school care provision increases full-time employment among mothers, but crowds out full-time employment among fathers. Thus, after-school care provision seems to contribute to the promotion of female labor supply and to an equal allocation of employment among parents.

Many developed countries consider an expansion of the childcare system. Besides care provision for preschool children, supplementary care for schoolchildren receives increasing attention. Switzerland, for example, has launched a federal program in

[64] Descriptive statistics for all three samples are shown in Tables 4.A.4 and 4.A.5 in the appendix.

2003, which provides subsidies to new or expanding care institutions. Germany is currently debating to extend its school system and to offer an increasing amount of all-day schools. Regarding maternal employment and female career opportunities, this investment might pay off: Our results indicate that each newly created after-school care slot causes one more mother to work full-time, as opposed to not working or working part-time. Yet, given the rather large confidence intervals of our estimates and the unknown general equilibrium effects, we abstain from providing a general policy recommendation.

Appendix

4.A Tables

Table 4.A.1: Cantonal involvement regarding childcare provision

Canton	Reference to childcare in cantonal legislation	Information/ coordination/ counseling	Reglementation*	Approval	Financial contribution
AG	No	Yes	No	No	Yes
AI	No	No	Yes	Yes	No
AR	Yes	No	No	No	No
BE	Yes	Yes	Yes	No	Yes
BL	Yes	Yes	No	No	Yes
BS	Yes	No	Yes	Yes	Yes
FR	Yes	Yes	Yes	Yes	No
GE	No	No	No	No	No
GL	No	No	Partially	Yes	Yes
GR	Yes	No	Yes	Yes	Yes
JU	Yes	Yes	Yes	Yes	Yes
LU	Yes	Yes	No	No	Yes
NE	Yes	Yes	Yes	Yes	Yes
NW	Yes	Yes	Yes	Yes	No
OW	Yes	Yes	Yes	Yes	Yes
SG	Yes	No	No	No	No
SH	No	No	Partially	No	No
SO	Yes	Yes	Yes	Yes	Yes
SZ	No	Yes	Partially	No	No
TG	Yes	Yes	Yes	Yes	No
TI	Yes	No	Yes	Yes	Yes
UR	No	Yes	No	No	Yes
VD	Yes	Yes	Yes	Yes	Yes
VS	Yes	Yes	Yes	No	Yes
ZG	Yes	Yes	Partially	No	No
ZH	Yes	Yes	Yes	No	No

Source: Internet platform Beruf und Familie (2008). Retrieved March 31, 2014, from http://www.berufundfamilie.admin.ch/ *Reglementation is under the responsibility of either the canton (= Yes), the municipality (= No) or both (= Partially).

Table 4.A.2: Cantonal school reforms and enforcement of after-school care provision

Canton	Year	Lunch care required by new school law	Afternoon care required by new school law	Enforcement of after-school care by 2010
	(1)	(2)	(3)	(4)
AG	2008	Yes	No	No
AI	-	-	-	No
AR	-	-	-	No
BE	2008	Yes	Yes	Yes
BL	2003	Yes	No	No
BS	2011	Yes	Yes	No
FR	-	-	-	No
GE	1997	Yes	Yes	Yes
GL	-	-	-	No
GR	2011	Yes	Yes	No
JU	-	-	-	No
LU	2012	Yes	Yes	No
NE	2015	No information	No information	No
NW	-	-	-	No
OW	-	-	-	No
SG	2008	Yes	No	No
SH	2012	Yes	Yes	No
SO	2007	Yes	Yes	Yes
SZ	-	-	-	No
TG	2005	No	No	No
TI	-	-	-	No
UR	-	-	-	No
VD	-	-	-	No
VS	-	-	-	No
ZG	-	-	-	No
ZH	2009	Yes	Yes	Yes

Source: Own investigations based on cantonal laws/school laws (2012). Abbreviations of cantons: ZH: Zürich, BE: Bern, LU: Luzern, UR: Uri, SZ: Schwyz, OW: Obwalden, NW: Nidwalden, GL: Glarus, ZG: Zug, FR: Fribourg, SO: Solothurn, BS: Basel Town, BL: Basel Country, SH: Schaffhausen, AR: Appenzell Ausserrhoden, AI: Appenzell Innerrhoden, SG: St. Gallen, GR: GraubÃijnden, AG: Aargau, TG: Thurgau,TI: Ticino, VD: Vaud, VS: Valais (Wallis), NE: Neuchâtel, GE: Geneva, JU: Jura.

Table 4.A.3: Treatment intensity and cut-off definition

Definition of cut-offs						
LLM	Above/ below cut-off	Obs.	Muni-cipalities	Cut-off (slots per child)	Average slots per child	Difference in slots per child
	(1)	(2)	(3)	(4)	(5)	(6)
1 BE-LU	Above	6848	22	0.004	0.052	0.052
	Below	7670	81		0	
2 ZH-LU	Above	6325	10	0.081	0.156	0.11
	Below	7439	17		0.047	
3 ZH-AG	Above	10577	43	0.05	0.106	0.081
	Below	10614	41		0.025	
4 ZH-AG	Above	12915	51	0.069	0.12	0.081
	Below	12799	49		0.039	
5 ZH-SH	Above	7138	36	0.068	0.145	0.11
	Below	7242	68		0.035	
6 ZH-TG	Above	7287	50	0.04	0.092	0.074
	Below	7495	51		0.018	
7 ZH-TG	Above	5695	24	0.04	0.084	0.07
	Below	8321	47		0.014	
8 ZH-SG	Above	2746	12	0.051	0.077	0.063
	Below	2861	20		0.015	
Total	Above	-	-	-	0.108	0.081
	Below	-	-	-	0.026	

Average after-school care (slots per child in the age 4–12) in municipalities above and below the LLM-specific cutoffs, by LLM. Calculation is based on all observations in the Swiss Structural Survey, sample restricted to individuals in the age 18–62. Abbreviations of cantons: AG: Aargau, BE: Bern, GR: Graubünden, TG: Thurgau, LU: Luzern, SG: St. Gallen, SH: Schaffhausen, ZH: Zürich.

Table 4.A.4: Descriptive statistics: Swiss women, age 18-62, w/ children age 0-12

	Pooled sample	Treated areas	Control areas	Difference	
	(1)	(2)	(3)	(4)	
	Mean	Mean	Mean	Diff.	p-val.
Labor Market Outcomes					
Employment (binary)	0.7	0.72	0.68	0.04	0.000
Full-time	0.1	0.11	0.09	0.01	0.017
Part-time	0.6	0.61	0.58	0.03	0.003
Low part-time	0.38	0.38	0.39	-0.01	0.322
Intermediate part-time	0.16	0.17	0.14	0.03	0.000
High part-time	0.05	0.06	0.05	0.01	0.024
Treatment/Instrument					
After-school care: Slots per child	0.06	0.11	0.03	0.08	0.000
Reform canton (binary)	0.32	0.56	0.28	0.28	0.000
Individual Control Variables					
Age	38.39	38.61	38.39	0.23	0.051
Mandatory education	0.09	0.1	0.08	0.02	0.002
Secondary education	0.55	0.53	0.58	-0.05	0.000
Tertiary education	0.35	0.36	0.33	0.03	0.003
Married	0.89	0.89	0.89	-0.01	0.405
Single	0.07	0.06	0.07	-0.01	0.260
Divorced	0.04	0.05	0.04	0.01	0.017
Widowed	0.01	0.01	0.01	0	0.455
Partner living in household	0.94	0.94	0.94	0	0.327
Number of kids	2.06	2	2.08	-0.07	0.000
Regional Control variables					
Vote share pro "Mutterschutz" (%)	0.45	0.48	0.43	0.05	0.000
Inhabitants	14925	18064	7123	10942	0.000
Urban	0.16	0.2	0.07	0.13	0.000
Agglomeration	0.46	0.57	0.53	0.05	0.000
Rural	0.38	0.23	0.41	-0.18	0.000
Income tax (100K; married & 2 kids,%)	6.62	6.3	6.56	-0.26	0.000
Population density/100 km2	795	932	599	332	0.000
Fraction of foreigners (%)	17.89	19.46	16.48	2.98	0.000
Unemployment rate	3.12	3.39	2.98	0.41	0.000
Home ownership in %	42	39	47	-8	0.000
Fraction of commuters (%)	59	61	63	-2	0.000

The sample is based on the Swiss Structural Survey 2010 (10,875 observations). Treated (control) areas are areas with a level of after-school care above (below) the cut-off.

Table 4.A.5: Descriptive statistics: Swiss men, age 18-62, w/ children age 0-12

	Pooled	Treated	Control	Difference	
	sample	areas	areas		
	(1)	(2)	(3)	(4)	
	Mean	Mean	Mean	Diff.	p-val.
Labor Market Outcomes					
Employment (binary)	0.97	0.97	0.98	-0.01	0.031
Full-time	0.89	0.88	0.91	-0.02	0.000
Part-time	0.08	0.09	0.07	0.02	0.004
Low part-time	0.03	0.03	0.02	0	0.195
Intermediate part-time	0.01	0.01	0.01	0	0.360
High part-time	0.04	0.04	0.03	0.01	0.015
Treatment/Instrument					
After-school care: Slots per child	0.06	0.11	0.03	0.08	0.000
Reform canton (binary)	0.32	0.56	0.29	0.27	0.000
Individual Control Variables					
Age	41.19	41.26	41.24	0.02	0.851
Mandatory education	0.05	0.05	0.04	0.01	0.029
Secondary education	0.39	0.37	0.4	-0.03	0.002
Tertiary education	0.54	0.56	0.54	0.02	0.024
Married	0.94	0.94	0.95	0	0.395
Single	0.04	0.04	0.04	0	0.451
Divorced	0.01	0.01	0.01	0	0.778
Widowed	0	0	0	0	0.123
Partner living in household	0.99	0.99	0.99	0	0.537
Number of kids	2.04	1.98	2.04	-0.06	0.000
Regional Control variables		.	.	.	.
Vote share pro "Mutterschutz" (%)	0.45	0.48	0.43	0.05	0.000
Inhabitants	14798	17926	7041	10884	0.000
Urban	787	919	594	325	0.000
Agglomeration	0.16	0.19	0.07	0.13	0.000
Rural	0.46	0.57	0.53	0.04	0.000
Income tax (100K; married & 2 kids,%)	0.38	0.23	0.41	-0.17	0.000
Population density/100 km2	6.62	6.3	6.56	-0.25	0.000
Fraction of foreigners (%)	17.75	19.23	16.4	2.83	0.000
Unemployment rate	3.1	3.36	2.97	0.39	0.000
Home ownership in %	42	40	47	-8	0.000
Fraction of commuters (%)	62	61	63	-2	0.000

The sample is based on the structural survey 2010 (10,133 observations). Treated (control) areas are areas with a level of after-school care above (below) the cut-off.

Table 4.A.6: Income taxes (canton and municipality component)

		# Munici- palities	Average tax	Median tax	Minimum tax	Maximum tax
	BE	50	8.8	8.8	8.1	9.6
1	LU	53	7.7	7.8	5.9	8.5
	Difference		1.1	1	2.1	1.2
	ZH	13	5.9	5.9	5.1	6.3
2	LU	14	7	7.3	5.2	7.5
	Difference		-1.2	-1.4	-0.2	-1.2
	ZH	24	5.8	5.9	4.9	6.3
3	AG	60	6.3	6.3	5.3	6.9
	Difference		-0.5	-0.4	-0.5	-0.6
	ZH	61	5.8	5.9	4.9	6.3
4	AG	40	6.1	6.2	5.3	6.9
	Difference		-0.4	-0.3	-0.5	-0.6
	ZH	79	6	6	4.9	6.3
5	SH	25	6.6	6.6	5.4	7.3
	Difference		-0.7	-0.6	-0.6	-1
	ZH	73	6	6	4.9	6.3
6	TG	28	6.9	7	6	7.6
	Difference		-0.9	-1	-1.1	-1.4
	ZH	22	6	6	5.4	6.3
7	TG	49	6.7	6.8	5.1	7.6
	Difference		-0.8	-0.8	0.3	-1.4
	ZH	22	6	6	5.4	6.3
8	SG	10	5.9	6	5	6.2
	Difference		0.1	0	0.3	0

Taxes are computed for individuals with 100,000 CHF income per year, married, with 2 kids. Taxes are reported in percentage points of total income. They include municipality taxes as well as cantonal taxes. Averages and median are unweighted. Abbreviations of cantons: AG: Aargau, BE: Bern, GR: Graubünden, TG: Thurgau, LU: Luzern, SG: St. Gallen, SH: Schaffhausen, ZH: Zürich.

Table 4.A.7: Preschool and primary school regulations across cantons (school year 2009/10)

LM - Canton	Minimum age at preschool entry	Preschool: hours/week (last preschool year)	Minimum age at school entry	Mandatory bloc hours (min 3.5 hours/working day)
1 - BE	4 yr. 3 m.	16.5 - 19.5	6 yr. 3 m.	yes
1- LU	4 yr. 9 m.	15 - 18	5 yr. 9 m.	yes
2 - ZH	4 yr. 3 m.	21 - 23	6 yr. 3 m.	yes
2 - LU	4 yr. 9 m.	15 - 18	5 yr. 9 m.	yes
3 - ZH	4 yr. 3 m.	21 - 23	6 yr. 3 m.	yes
3 - AG	4 yr. 3 m.	21 - 25	6 yr. 3 m.	no
4 - ZH	4 yr. 3 m.	21 - 23	6 yr. 3 m.	yes
4 - AG	4 yr. 3 m.	21 - 25	6 yr. 3 m.	no
5 - ZH	4 yr. 3 m.	21 - 23	6 yr. 3 m.	yes
5 - AG	4 yr. 3 m.	20.4	6 yr. 3 m.	yes
6 - ZH	4 yr. 3 m.	21 - 23	6 yr. 3 m.	yes
6 - TG	4 yr. 3 m.	21 - 25	6 yr.	no**
7 - ZH	4 yr. 3 m.	21 - 23	6 yr. 3 m.	yes
7 - TG	4 yr. 3 m.	21 - 25	6 yr.	yes
8 - ZH	4 yr. 3 m.	21 - 23	6 yr. 3 m.	yes
8 - SG	4 yr.	24	6 yr.	yes

Continued on the next page.

Continued from the previous page: Preschool and primary school regulations across cantons (school year 2009/10)

LM - Canton	Preschool: Mandatory offer by municipality in years	Attendance: Fraction of children with 1 preschool year*	Attendance: Fraction of children with 2 preschool years
1 - BE	1	19%	80%
1- LU	1	63%	37%
2 - ZH	2	2.20%	95.70%
2 - LU	1	63%	37%
3 - ZH	2	2.20%	95.70%
3 - AG	1	2%	96%
4 - ZH	2	2.20%	95.70%
4 - AG	1	2%	96%
5 - ZH	2	2.20%	95.70%
5 - AG	2	1.80%	98%
6 - ZH	2	2.20%	95.70%
6 - TG	2	1%	96%
7 - ZH	2	2.20%	95.70%
7 - TG	2	1%	96%
8 - ZH	2	2.20%	95.70%
8 - SG	2	ca. 10%	ca. 90%

*Fraction is computed with respect to all children in their first year in primary school. **Introduced: 2010 - 2013. Abbreviations of cantons: AG: Aargau, BE: Bern, GR: Graubünden, TG: Thurgau, LU: Luzern, SG: St. Gallen, SH: Schaffhausen, ZH: Zürich

Table 4.A.8: Distance to economic hubs (avg. commuting times by car in minutes)

LLM	Canton	Capital of canton with childcare regulation	Capital of canton without childcare regulation
		Berne	Luzern
	BE	46	52
1	LU	71	29
	Difference	-25	23
		Zurich	Luzern
	ZH	22	32
2	LU	41	16
	Difference	-19	16
		Zurich	Aarau
	ZH	21	27
3	AG	31	15
	Difference	-10	12
		Zurich	Aarau
	ZH	24	26
4	AG	29	19
	Difference	-5	7
		Zurich	Schaffhausen
	ZH	34	30
5	SH	52	18
	Difference	-18	12
		Zurich	Frauenfeld
	ZH	34	45
6	TG	49	25
	Difference	-15	20
		Zurich	Frauenfeld
	ZH	36	38
7	TG	52	22
	Difference	-17	16
		Zurich	St. Gallen
	ZH	36	43
8	SG	52	25
	Difference	-16	18

The upper canton in each panel is the canton with cantonal childcare regulation; the lower canton is the canton without childcare regulation. Only municipalities in LLMs are included. The displayed distances correspond to unweighted averages over municipalities in each of the canton. Abbreviations of cantons: AG: Aargau, BE: Bern, GR: Graubünden, TG: Thurgau, LU: Luzern, SG: St. Gallen, SH: Schaffhausen, ZH: Zürich.

Table 4.A.9: Placebo estimation using men and women age 18-42 without children

	Childcare above the cut-off	Childcare below the cut-off	Effect	95 % CI	
	(1)	(2)	(3)	(4)	
Panel A) Swiss Women without children (age <42)					
Employment (binary)	0.76	0.75	0.01	-0.07	0.15
Full-time (binary)	0.55	0.51	0.04	-0.06	0.22
Part-time (binary; < 36h/week)	0.21	0.24	-0.04	-0.13	0.06
Low part-time (binary; < 20h/week)	0.06	0.07	-0.01	-0.08	0.03
Intermediate part-time (binary; 20-27h/week)	0.05	0.05	0	-0.07	0.04
High part-time (binary; 28-35h/week)	0.1	0.12	-0.02	-0.06	0.08
Effect of Reform on Child Care Supply (First Stage)	0.72	0.3	0.42***	0.29	0.57
Panel B) Swiss Men without children (age <42)					
Employment (binary)	0.78	0.78	-0.01	-0.15	0.12
Full-time (binary)	0.67	0.71	-0.04	-0.19	0.09
Part-time (binary; < 36h/week)	0.1	0.07	0.03	-0.06	0.11
Low part-time (binary; < 20h/week)	0.05	0.02	0.03	-0.04	0.08
Intermediate part-time (binary; 20-27h/week)	0.02	0.02	0	-0.05	0.03
High part-time (binary; 28-35h/week)	0.03	0.03	0	-0.03	0.07
Effect of Reform on Child Care Supply (First Stage)	0.69	0.29	0.39***	0.25	0.57

*significant at the 1%; **significant at the 5%; *significant at the 10%-level. Estimates are weighted averages of the instrumental variable estimates for each LLM. The underlying weights correspond to the number of compliers in the respective LLM. The instrument is based on the enforcement of after-school care supply in the cantonal school law. The sample corresponds to 16,381 observations in the case of women and 18,652 observations in the case of men.

Table 4.A.10: Local childcare supply prior to and after changing the municipality of residence

Coverage with after-school care (slots per child), before and after moving to a new municipality						
Panel A: Female (age 18–62)						
	Municipality inside of LLM area after change of residence			Municipality inside of LLM area before change of residence		
Age	Observations	Coverage rate (Slots per child)		Observations	Coverage rate (Slots per child)	
		Before	After		Before	After
(1)	(2)	(3)	(4)	(5)	(6)	(7)
18-22	289	0.07	0.07	285	0.06	0.09
23-27	470	0.08	0.07	504	0.07	0.1
28-32	429	0.1	0.07	431	0.07	0.1
33-37	295	0.11	0.06	267	0.08	0.08
38-42	189	0.09	0.06	195	0.07	0.09
43-47	164	0.09	0.07	170	0.07	0.08
48-52	130	0.08	0.06	131	0.07	0.07
53-57	86	0.09	0.06	94	0.08	0.08
58-62	53	0.1	0.06	55	0.07	0.07

Continued on the next page.

Continued from the previous page: Local childcare supply prior to and after changing the municipality of residence

Coverage with after-school care (slots per child), before and after moving to a new municipality						
Panel B: Male (age 18–62)						
	Municipality inside of LLM area after change of residence			Municipality inside of LLM area before change of residence		
Age	Observations	Coverage rate (Slots per child)		Observations	Coverage rate (Slots per child)	
		Before	After		Before	After
(1)	(2)	(3)	(4)	(5)	(6)	(7)
18-22	202	0.07	0.07	207	0.06	0.09
23-27	402	0.07	0.07	411	0.06	0.1
28-32	437	0.1	0.07	429	0.08	0.1
33-37	304	0.1	0.07	316	0.07	0.1
38-42	254	0.11	0.07	232	0.08	0.09
43-47	185	0.09	0.07	182	0.08	0.08
48-52	136	0.11	0.07	125	0.07	0.08
53-57	90	0.09	0.06	99	0.06	0.08
58-62	77	0.11	0.06	72	0.08	0.07

Sample based on Swiss Structural Survey. All individuals age 18–62 who have migrated between two municipalities within the last 12 months before the survey are included. Columns 2-4: Individuals who are living inside the LLM area after migration. To make sure we do not neglect individuals who leave the area covered by the LLMs, Columns 5-7 refer to individuals who have been living in the LLM area before migration. The two samples are overlapping, i.e. both contain individuals who have been living in the LLM area both before and after migrating to a new municipality.

Table 4.A.11: Propensity score estimations for each local labor market separately

	LLM 1		LLM 3		LLM 4		LLM 5		LLM 6		LLM 8	
Low education	0.009		0.451	***	0.3	***	0.046		-0.123		0.023	
High education	-0.217	**	0.087		-0.001		-0.146		-0.051		-0.066	
Age	-0.046		-0.007		0.063		0.128	*	0.19	***	0.11	
Age squared	0		0		-0.001		-0.001		-0.002	***	-0.001	
Partner	-0.273		0.05		-0.027		0.01		0.267	*	0.152	
Local referendum results	9.377	***	6.253	***	-3.333	***	3.725	***	11.196	***	13.427	***
# children	-0.093		-0.038		0.022		-0.207	***	-0.176	***	-0.129	*
# children age 0–4	-0.054		0.008		-0.1	**	0.176	**	0.072		0.054	
# children age 5–12	0.114		0.04		-0.022		0.145	**	0.127	**	0.129	
Constant	-3.131	**	-3.618	***	0.513		-3.359	**	-8.505	***	-8.903	***
Observations	1440		1595		2050		1303		1385		1177	

This table displays the coefficients resulting from the propensity score estimation. In other words, these coefficients stem from a probit estimation of the instrument (binary variable indicating whether the individual lives in a canton enforcing after-school care provision) on the set of control variables listed above. Only LLMs with a positive share of compliers. *Significant at the 1%; ** significant at the 5%; *significant at the 10% significance level.

Table 4.A.12: External validity: Descriptive statistics, pooled sample versus complier sample

	Pooled Sample		Complier Sample	
	Mean	Std. error	Mean	Std. error
Panel A: Women, 18-62 years old, with children 0-12 years old				
Labor Market Outcomes				
Employment (binary)	0.70	0.00	0.76	0.06
Full-time	0.10	0.00	0.12	0.05
Part-time	0.60	0.00	0.64	0.07
Low part-time	0.38	0.00	0.45	0.07
Intermediate part-time	0.16	0.00	0.14	0.07
High part-time	0.05	0.00	0.06	0.04
Individual Control Variables				
Age	38.39	0.06	38.06	0.05
Mandatory education	0.09	0.00	0.1	0.04
Tertiary education	0.35	0.00	0.37	0.07
Partner living in household	0.94	0.00	0.94	0.03
Number of kids	2.06	0.01	2.01	0.09
Regional Control variables				
Vote share pro "Mutterschutz" (%)	0.45	0.00	0.47	0
Panel B: Men, 18-62 years old, with children 0-12 years old				
Labor Market Outcomes				
Employment (binary)	0.97	0.00	0.94	0.04
Full-time	0.89	0.00	0.87	0.06
Part-time	0.08	0.00	0.07	0.05
Low part-time	0.03	0.00	0.03	0.03
Intermediate part-time	0.01	0.00	0.01	0.02
High part-time	0.04	0.00	0.04	0.03
Individual Control Variables				
Age	41.19	0.06	40.91	0.04
Mandatory education	0.05	0.00	0.05	0.04
Tertiary education	0.54	0.00	0.53	0.07
Number of kids	2.04	0.01	1.99	0.13
Regional Control variables				
Vote share pro "Mutterschutz" (%)	0.45	0.00	0.46	0

The pooled sample provides the unweighted descriptive statistics for all observations included in the LLMs. The sample corresponds to 10,875 observations in the case of women and 10,133 observations in the case of men. The complier sample provides the moments calculated using the weighted averages of the IV estimates for each LLM (LATE). The underlying weights correspond to the number of compliers in the respective LLM. The instrument is based on the enforcement of after-school care supply in the cantonal law.

Table 4.A.13: External Validity: Men and women, age 18-62, w/ at least one child age 0-12

	LLM	German- speaking CH	LLM - German- speaking CH	
	Mean	Mean	Diff.	p-val.
Labor Market Outcomes				
Employment (binary)	0.83	0.84	0	0.213
Full-time	0.48	0.47	0.01	0.127
Part-time	0.35	0.36	-0.01	0.011
Low part-time	0.21	0.2	0.01	0.006
Intermediate part-time	0.09	0.1	-0.01	0
High part-time	0.05	0.06	-0.01	0
Treatment/Instrument				
After-school care: Slots per child	0.06	0.09	-0.03	0
Reform canton (binary)	0.32	0.43	-0.11	0
Individual Control Variables				
Age	39.74	39.79	-0.05	0.427
Female	0.52	0.52	0	0.884
Mandatory education	0.07	0.08	-0.01	0.003
Secondary education	0.48	0.44	0.03	0
Tertiary education	0.44	0.47	-0.02	0
Married	0.91	0.91	0	0.983
Single	0.06	0.06	0	0.739
Divorced	0.03	0.03	0	0.835
Widowed	0	0	0	0.062
Partner living in household	0.97	0.96	0	0.036
Number of kids	2.05	2.04	0.01	0.112
Regional Control variables				
Vote share pro "Mutterschutz"	0.45	0.48	-0.03	0
No. of inhabitants in 2010	14864	66742	-51878	0
Population density per 100 km2	791	1340	-548	0
Urban	0.16	0.31	-0.15	0
Agglomeration	0.46	0.39	0.06	0
Rural	0.38	0.3	0.08	0
Income tax at 100K married & 2 kids (%)	6.62	6.61	0.01	0.432
Unemployment rate	3.11	3.08	0.03	0.024
Home ownership in %	42	37	5	0
Fraction of commuters (%)	59	51	9	0

Sample: German language region. Males and females in the age 18-62, with at least one child in the age of 0-12 (n = 46,428). 13,775 individuals living inside an LLM, 32,653 individuals living outside an LLM.

Table 4.A.14: Weighting scheme based on the share of compliers: Results for men and women with children (age 0-12)

	Municipalitites with cantonal enforcement of after-school care	Municipalitites without cantonal enforcement of after-school care	Effect	95% CI	
	(1)	(2)	(3)	(4)	
Panel A) Swiss Women with children (age 0-12)					
Employment (binary)	0.76	0.68	0.08	-0.04	0.21
Full-time (binary)	0.12	0.03	0.09*	0	0.19
Part-time (binary; < 36h/week)	0.64	0.65	-0.01	-0.14	0.13
Low part-time (binary; < 20h/week)	0.45	0.43	0.02	-0.13	0.16
Intermediate part-time (binary; 20-27h/week)	0.14	0.16	-0.03	-0.19	0.09
High part-time (binary; 28-35h/week)	0.06	0.06	0	-0.06	0.09
Effect of Instrument on treatment					
Panel B) Swiss Men with children (age 0-12)					
Employment (binary)	0.94	0.96	-0.02	-0.1	0.03
Full-time (binary)	0.87	0.97	-0.09*	-0.22	0.02
Part-time (binary; < 36h/week)	0.07	-0.01	0.07	-0.02	0.18
Low part-time (binary; < 20h/week)	0.03	0	0.02*	-0.01	0.09
Intermediate part-time (binary; 20-27h/week)	0.01	-0.02	0.02*	-0.01	0.06
High part-time (binary; 28-35h/week)	0.04	0.01	0.03	-0.06	0.08
Effect of Instrument on treatment					

*significant at the 1%; **significant at the 5%; *significant at the 10%-level. Above estimates are weighted averages of the instrumental variable estimates for each LLM. The underlying weights correspond to the share of compliers in the respective LLM. The instrument is based on the enforcement of after-school care supply in the cantonal school law. The sample corresponds to 10875 observations in the case of women and 10133 observations in the case of men.

Table 4.A.15: Weighting scheme based on the number of observations: Results for men and women with children (age 0-12)

	Municipalitites with cantonal enforcement of after-school care	Municipalitites without cantonal enforcement of after-school care	Effect	95% CI	
	(1)	(2)	(3)	(4)	
Panel A) Swiss Women with children (age 0–12)					
Employment (binary)	0.76	0.71	0.05	-0.41	0.58
Full-time (binary)	0.12	0.01	0.11	-0.35	0.56
Part-time (binary; < 36h/week)	0.65	0.7	-0.05	-0.55	0.54
Low part-time (binary; < 20h/week)	0.45	0.46	0	-0.53	0.76
Intermediate part-time (binary; 20-27h/week)	0.14	0.17	-0.03	-0.65	0.47
High part-time (binary; 28-35h/week)	0.06	0.07	-0.01	-0.39	0.35
Effect of Instrument on treatment	0.56	0.3	0.26	0.11	0.35
Panel B) Swiss Men with children (age 0–12)					
Employment (binary)	0.95	0.97	-0.02	-0.2	0.13
Full-time (binary)	0.88	0.97	-0.1	-0.76	0.55
Part-time (binary; < 36h/week)	0.07	-0.01	0.08	-0.46	0.7
sLow part-time (binary; < 20h/week)	0.02	0	0.02	-0.16	0.25
Intermediate part-time (binary; 20-27h/week)	0.01	-0.01	0.02	-0.12	0.14
High part-time (binary; 28-35h/week)	0.04	0.01	0.03	-0.27	0.42
Effect of Instrument on treatment	0.55	0.32	0.24***	0.07	0.35

*significant at the 1%; **significant at the 5%; *significant at the 10%-level. Above estimates are weighted averages of the instrumental variable estimates for each LLM. The underlying weights correspond to the number of observations in the respective LLM. The instrument is based on the enforcement of after-school care supply in the cantonal school law. The sample corresponds to 10,875 observations in the case of women and 10,133 observations in the case of men.

Table 4.A.16: Weighting scheme based on the number of observations with zero weights given to LLMs with negative first stage results: Results for men and women with children (age 0–12)

	Municipalitites with cantonal enforcement of after-school care	Municipalitites without cantonal enforcement of after-school care	Effect	95% CI	
	(1)	(2)	(3)	(4)	
Panel A) Swiss Women with children (age 0–12)					
Employment (binary)	0.78	0.69	0.09	-0.13	0.56
Full-time (binary)	0.12	0.01	0.12*	-0.02	0.62
Part-time (binary; < 36h/week)	0.65	0.69	-0.03	-0.39	0.26
Low part-time (binary; < 20h/week)	0.45	0.46	-0.01	-0.39	0.4
Intermediate part-time (binary; 20-27h/week)	0.13	0.16	-0.03	-0.6	0.17
High part-time (binary; 28-35h/week)	0.07	0.06	0.01	-0.11	0.39
Effect of Instrument on treatment	0.63	0.25	0.39***	0.25	0.51
Panel B) Swiss Men with children (age 0–12)					
Employment (binary)	0.95	0.97	-0.03	-0.24	0.03
Full-time (binary)	0.88	1	-0.12*	-0.95	0
Part-time (binary; < 36h/week)	0.07	-0.03	0.1	-0.06	0.77
Low part-time (binary; < 20h/week)	0.03	0	0.03	-0.04	0.27
Intermediate part-time (binary; 20-27h/week)	0.01	-0.02	0.03	-0.03	0.18
High part-time (binary; 28-35h/week)	0.04	0	0.04	-0.06	0.44
Effect of Instrument on treatment	0.63	0.27	0.36***	0.24	0.49

*significant at the 1%; **significant at the 5%; *significant at the 10%-level. Above estimates are weighted averages of the instrumental variable estimates for each LLM. The underlying weights correspond to the number of observations in the respective LLM with zero weights given to defiers. The instrument is based on the enforcement of after-school care supply in the cantonal school law. The sample corresponds to 10,875 observations in the case of women and 10,133 observations in the case of men.

4.B Construction of local labor markets

To construct local labor markets (LLMs), we draw upon the 160 Swiss "Mobilité Spatiale regions" (henceforth "MS regions"), which were defined in 1982 by the Statistical office of Switzerland based on commuting behavior. We combine all MS regions that lie within a limited commuting area (30 minutes by car) and that lie along a cantonal border that signifies a division in the cantonal regulation of after-school care services.[65] We drop all LLMs i) where the area on one side of the cantonal border contains the majority of the respective cantonal population;[66] ii) where the populations to both sides of the cantonal border differ strongly in their preferences related to work and family; and iii) where there is no clear division in the preferences related to work and family between the municipalities inside and outside the LLM in at least one of the two cantons considered in the respective LLM (see the empirical evidence below).

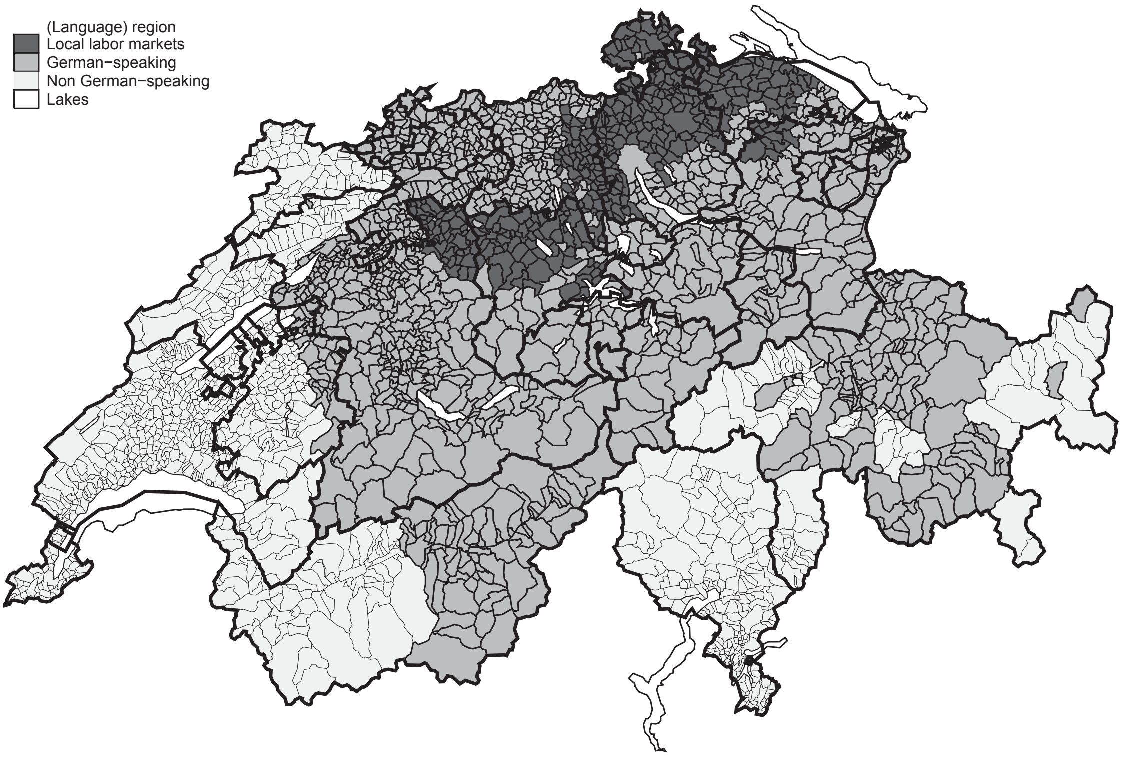
The resulting LLMs are either municipalities at the cantonal division of Bern with the surrounding cantons (here Lucerne) or municipalities at the cantonal division of Zurich with the surrounding cantons (here Aargau, Lucerne, Schaffhausen, St. Gallen and Thurgau).[67] Figure 4.B represents the geographical area covered by the LLMs. Although the geographical area is rather small, it contains 20% of the overall Swiss population (and 30% of the overall Swiss German population).

[65] Note that LLMs can overlap. Yet, we only consider LLMs that contain exactly one cantonal border, i.e. that contain municipalities from exactly two different cantons.

[66] We deviate twice from this condition, in LLM 5 and in LLM 7. Yet, the discontinuity regarding the cantonal legislation and thus the after-school care provision across the cantonal border is in both cases driven by the other cantonal part. In other words, there is at least one cantonal part where the population living inside the LLM is outvoted by the population living outside the LLM.

[67] There are two further potential sets of cantonal borders: borders of the canton Solothurn and its neighbor cantons, and borders between the cantons Geneva and Vaud. Because of the lack of data on after-school care for Solothurn, we cannot use any LLM based on Solothurn and the neighboring cantons. The LLM along the cantonal border between Geneva and Vaud cannot be used for our analysis, as there is no strong heterogeneity in the preferences regarding work and family within the respective cantons. One further potential LLM stretching over the cantonal border between Zurich and Zug is excluded as income taxes, an issue discussed below, are substantially different in both cantons.

Figure 4.B.1: Geographical area covered by LLMs



Source: Own calculations.

Table 4.1 in the main text lists the resulting LLMs. Bern and Zurich are cantons that by 2010 (the year of our data) explicitly enforce after-school care – and thus observations belonging to these cantons are assigned the value one for the IV. The remaining cantons Aargau, Lucerne, Schaffhausen, St. Gallen and Thurgau, do not explicitly enforce after-school care in their cantonal legislation by 2010 – and thus observations belonging to these cantons are assigned the value zero for the IV.

Table 4.1, column 6, provides descriptive evidence for the cantonal borders to be monotone and strong IVs. Cantonal laws enforcing after-school care supply indeed correlate positively with after-school care provision. With the exception of one LLM, there is a higher supply of after-school care in the municipalities of the canton legally enforcing after-school care provision than in the municipalities of the canton not legally enforcing after-school care provision.[68]

Table 4.1, columns 4 and 5, provide some supportive evidence that the cantonal school law is exogenous to the preferences related to work and family of the population residing in municipalities within the LLM. First, the municipalities included in the LLMs correspond on at least one side of the cantonal border to the minority of the respective cantonal population. Second, the populations to both sides of the cantonal border share the same preferences regarding work and family. To address this issue, we rely on the results of the referendum on maternity benefits (September 26, 2004). Results on the referendum are rather similar across the cantonal border within each LLM. Yet, on at least one side of the cantonal border, the remaining cantonal population outside the LLM outvotes the population living inside the LLM.

Using the example of the LLM along the cantonal border between Bern and Lucerne helps to illustrate this issue. Inside the LLM the referendum failed to both sides of the cantonal border. It also failed in the remaining municipalities of the canton Lucerne. Yet, the respective municipalities belonging to the canton Bern were outvoted by the remaining cantonal population. Hence, while citizens inside the LLM are rather similar regarding their preferences related to work and family,

[68] When aggregating the estimates for the different LLMs, we weight each estimate by the number of compliers inside the respective LLM and thus, any defiers – municipalities that decrease their after-school care because of the legal enforcement – are not taken into consideration.

the remaining cantonal population outside the LLM differs, in at least one of the two cantons, strongly with respect to such preferences. As a result, differences in the existing cantonal laws related to work and family might arise, but are unlikely to be driven by the population living in the municipalities belonging to the LLM.

5. Financial Work Incentives for Disability Benefit Recipients: Lessons from a Randomized Field Experiment

Monika Bütler, Eva Deuchert, Michael Lechner, Stefan Staubli, and Petra Thiemann

Abstract

Disability insurance (DI) beneficiaries lose part of their benefits if their earnings exceed certain thresholds (“cash-cliffs”). This implicit taxation is considered the prime reason for low DI outflow. We analyze a conditional cash program that incentivizes work related reductions of disability benefits in Switzerland. 4,000 randomly selected DI recipients receive an offer to claim up to CHF 72,000 (USD 71,000) if they expand work hours and reduce benefits. Initial reactions to the program announcement, measured by call-back rates, are modest; individuals at cash-cliffs react more frequently. By the end of the field phase, the take-up rate amounts to only 0.5%.

JEL codes: H55, J14, C93, D04
Keywords: disability insurance, field experiment, financial incentives, return-to-work

5.1 Introduction

The high number of disability insurance (DI) recipients – about 6% of the working-age population of OECD countries received disability benefits in 2007 – generates high costs to society. In 2007, OECD countries spent on average 1.2% of their GDP on DI benefits, which is almost 2.5 times higher than the fraction of GDP spent on unemployment benefits. Outflow from DI receipt is low at 1–2% per year (OECD, 2003, 2009, 2010).[69] Work disincentives are considered a major cause for the low outflow from DI (OECD, 2010): In most countries, DI recipients lose (part of) their benefits if their earnings increase beyond certain thresholds ("cash-cliffs"). Therefore, the OECD advocates reforms that increase return-to-work incentives.

Empirical evidence on the effectiveness of such reforms however is scarce.[70] Campolieti and Riddell (2012) evaluate a change in the "earnings disregard", which is the amount of earnings that DI recipients are allowed to receive without losing their benefits; Kostøl and Mogstad (2014) as well as Weathers and Hemmeter (2011) investigate the introduction of financial work incentives; and Gettens (2009) analyses the effect of expanding health insurance coverage to individuals who exit from DI into employment. While some of these policies increased employment, none of them affected DI outflow.[71] To our knowledge, no study so far examines conditional cash incentives that are paid out if individuals reduce their benefits or even exit the DI.

This paper complements the literature with results of a field experiment in Switzerland: To stimulate employment and benefit reduction, the DI offered a conditional cash transfer ("seed capital") to 4,000 randomly selected DI recipients. The seed capital program differs in two ways from previous programs: First, eligibility de-

[69] Here, we do not count outflow into retirement.

[70] Other types of DI reforms include policies that reduce DI inflow, such as reducing benefit generosity, altering eligibility criteria, or implementing stricter screening. These policies are relatively successful in reducing the number of DI recipients (de Jong et al., 2011, Staubli, 2011, Low and Pistaferri, 2010, Vuren and Vuuren, 2007). Policies that aim at increasing DI outflow by providing access to vocational rehabilitation and employment integration are less effective. Results indicate low take-up and no or only small effects on outflow (Adam et al., 2010, Stapleton et al., 2008, Thornton et al., 2004, Kornfeld and Rupp, 2000).

[71] The Medicaid expansion described by Gettens (2009) had no employment effects.

pends directly on employment outcomes and benefit reduction. Individuals can only claim seed capital if they take up or expand employment, and if, as a consequence, their disability pension decreases by at least one quarter.[72] Second, the financial incentive is large compared to incentives in previously studied programs. Individuals receive a one-time payment of 18,000 Swiss francs (CHF) in the high treatment condition, or CHF 9,000 in the low treatment condition for a reduction of disability benefits by one quarter. The maximum payment to an individual with a full pension who completely exits the DI thus amounts to CHF 72,000 (about USD 71,000 at the time of the introduction of the program in September 2010). This amount compares to the average disposable yearly income of Swiss households (FSO, 2004). In addition, the lump-sum payment does not depend on the benefit level and enjoys preferential tax treatment.

The program did not succeed in increasing outflow and employment. By the end of the field phase (September 2010-August 2013), only 0.5% of individuals took up seed capital. This number is approximately as high as the average number of pension reductions in previous years (0.4%). Thus, seed capital supposedly generated windfall gains for some individuals rather than true work incentives. Furthermore, we evaluate case worker contacts within five months after individuals received a seed capital offer letter. Despite encouragement, only 4% of individuals contacted their local case worker for more information; offering a higher payment did not change this response pattern.

This paper attributes the low take-up primarily to an insufficient size of the incentives. We present micro-simulation results, based on rich income and employment information from survey and administrative data, which both cover the pre-program period. For a majority of individuals, extending labor supply for a period of more than two years would not have been beneficial. The simulations predict higher take-up rates only for individuals with particularly strong work disincentives. These are individuals who would lose a substantial amount of their benefits, but who would

[72] A reduction in DI benefits is thus driven by an increase in labor supply. This is in contrast to papers that study the labor supply response to a change in DI benefits (Autor and Duggan, 2007, Marie and Vall Castello, 2012).

gain only little in terms of earnings, if they extended their labor supply beyond a certain threshold (“cash-cliff”). Indeed, individuals close to cash-cliffs are more likely to contact their case worker in the initial program phase, but the magnitude of this effect is small.

We discuss three further reasons for the low take-up: risk aversion (i.e., individuals hesitate to trade safe income from benefits against risky income from wages), bounded rationality (i.e., case workers apply rules of thumb to assess an individual’s disability degree, which undermines our classification of cash-cliff constrained individuals), and information frictions (i.e., individuals did not read or did not process the offer letter).

The paper proceeds as follows: Section 5.2 provides a description of the disability insurance system in Switzerland and discusses the design of the experiment. Section 5.3 describes the data. Section 5.4 outlines the expected impact in a standard labor supply model and presents simulation results of the program effects. Section 5.5 summarizes the results, followed by a discussion in Section 5.6. Section 5.7 concludes.

5.2 The Swiss disability insurance system and the experiment

5.2.1 An overview of the institutional setting

In Switzerland, individuals who partially or fully lose their ability to work due to health impairments can claim disability benefits. These benefits come from three different social security programs:[73] First, the mandatory public disability insurance serves all persons who live or work in Switzerland (“first pillar”). Second, the mandatory employer-based occupational pension scheme applies to all employees whose annual earnings exceed CHF 20,000 (“second pillar”). Third, the supplemen-

[73] If the disability is caused by an accident or an occupational disease, then it is likely that pensions are paid from the public accident insurance scheme. This insurance type, however, is not the focus of this paper and is thus not further considered.

tary benefit scheme grants means-tested benefits to individuals in need. These are individuals who cannot cover basic costs of living with the benefits from the first two pillars as well as with other income sources (comparable to the Supplemental Security Income in the US). The generosity of these three different programs depends on various factors, such as contribution years, average lifetime earnings, and the number of dependent children. The first two pillars guarantee a replacement rate of 60-80% (net of tax). Means-tested benefits secure an income of CHF 3,000 for singles and CHF 4,500 for couples, in addition to health care costs (see Figure 5.A.1 for an example of a benefit pattern).

Individuals who only partially lose their ability to work are eligible for a "partial" pension (first and second pillar); many DI recipients thus work at least part-time (37%, see Table 5.A.2). The amount of the partial pension depends on an individuals' DI degree, which is his/her hypothetical earnings loss due to disability.[74] DI recipients receive a quarter pension with a disability degree between 40% and 49%, a semi pension with a disability degree between 50% and 59%, a three-quarter pension with a disability degree between 60% and 69%, and a full pension with a disability degree of 70% and higher. Thus, whereas the disability degree is a continuous function of the earnings loss, the pension is a step-function of the earnings loss.

To calculate the disability degree, DI case workers assess two types of potential earnings: "potential earnings without disability" and "potential earnings with disability". They typically predict the former based on an individual's earnings before disability. Similarly, they typically evaluate the latter based on an individual's earnings during disability. This procedure is valid only if the DI beneficiary always exhausts his or her (remaining) work capacity. Therefore, if the case worker concludes that the insured person does not fully exhaust his or her work capacity, he can fix potential earnings based on assumed work capacity and based on official wage indices.

[74] Partial DI systems are known in many countries (such as Norway, the Netherlands, Sweden, or Germany for example). The decision to award a full or a partial DI pensions is, however, typically based on functional limitations or the number of hours a person can perform in a job rather than on potential earnings.

In practice, this assessment procedure is likely to be imperfect because of the lack of objective information on work capacity. On the one hand, case workers might use rules of thumb and thus award certain salient disability degrees more often (e.g., 50%). On the other hand, DI recipients can signal low work capacity by not taking up a job or by working only a small number of hours. They might thus influence their disability degree and, consequently, the size of their disability pension. The stepwise benefit structure potentially reinforces this asymmetric information problem. In order to maintain higher benefit levels, individuals might choose low working hours, even if they recover from their disability. Incentives to signal high DI degrees through low working hours are particularly strong for certain subgroups (see further explanations in Section 5.4). The field experiment described in this paper tests one potential avenue to reduce these work disincentives.

5.2.2 Experiment "Pilot Project Seed Capital"

To measure the effect of a reduction in financial work disincentives for DI recipients on DI outflow, we conducted a field experiment ("Pilot Project Seed Capital") in collaboration with the Swiss Federal Social Insurance Office (henceforth "FSIO"). Seed capital is a conditional lump-sum payment for DI recipients who meet two requirements: First, they have to take up or expand work in the primary labor market, which should lead to an earnings increase. Second, this earnings increase has to be large enough to trigger a pension reduction by at least one quarter (e.g., from a semi pension to a quarter pension). A fall-back rule accommodates a potential deterioration in health status: Within five years after the pension reduction, individuals can fall back to their old DI contract if they cannot work for 30 consecutive days (see Section 5.4).

To test different amounts of the financial incentives, we implemented two treatment conditions, that is, "high" seed capital (CHF 18,000 per pension reduction by one quarter), and "low" seed capital (CHF 9,000 per pension reduction by one quarter). Thus, the maximally achievable seed capital, that is, the seed capital for a person with a full pension who completely exits the DI, amounts to either CHF 72,000 (high seed capital) or CHF 36,000 (low seed capital). Whereas the former amount compares to the average income of a Swiss household, the latter amount corresponds to a minimum yearly income, which is implicitly guaranteed by means-tested benefits. The DI splits the lump-sum payment in four equal tranches, paid bi-annually over two years. Once an individual falls back to a higher pension, the DI stops the payment of outstanding seed capital tranches.

Two cantons participated in the field experiment, St. Gallen, a German speaking canton, and Vaud, a French speaking canton. Out of the 37,853 DI recipients in these two cantons, we randomly chose 6,020 individuals for the two treatment conditions (2,000 individuals each) and for the control condition (2,020 individuals). Table 5.A.1 provides details on the stratified assignment mechanism.

The field phase of the experiment took place between September 2010 and August 2013. In September 2010, a letter from the local DI offices informed the treated individuals about seed capital eligibility. This letter explained the eligibility rules as well as the fall-back rule mentioned above. Furthermore, the letter encouraged participants to contact their DI case worker for further information and assistance (see the complete letter in Section 5.D). The control group did not receive any information. After contacting the DI office by phone, individuals could meet their DI case workers in person to discuss further integration steps.

5.3 Data

For both the design and the evaluation of the program, we use three different data sources: administrative data from the Swiss pension system, baseline survey data, and case worker records on individual contacts with DI recipients.

To choose program participants and to simulate program effects, we combine administrative data from the Swiss pension system (first pillar) with baseline survey data. Both datasets cover the pre-program period. The administrative data include all DI recipients in the participating cantons, and contain full labour market histories, demographic characteristics, and information on first-pillar pensions, but not, however, data on further income sources (such as second pillar and means-tested benefits). To enrich the administrative database, we thus conducted a telephone survey among 8,000 randomly selected individuals prior to program announcement (response rate: 51%). The survey data capture current employment, detailed information on all possible income sources (i.e., wages, work hours, second pillar benefits, means-tested benefits, partners' income), further demographic characteristics (e.g., marital status, number of children, and education), and information on work capacity (e.g., health status, perceived difficulty to find a job).

To assess the program response in the short-run, we match the above data with case worker records on all interactions with individuals in the treatment groups, starting at the time of the program announcement.[75] The data consist of the date, the frequency, and the content of all interactions that took place both over the phone and in person, for up to five months after the program announcement (i.e., between September 2010 and February 2011).

As documented interest during the first five months of the experiment fell far behind the FSIO's expectations, the FSIO refrained from further data collection. Furthermore, a low take-up rate of only 0.5%, that is, 20 individuals, prevents further quantitative investigations into long-term outcomes. Unfortunately, for data security reasons, we were also unable to collect further post-treatment survey data that would allow us to deeply assess the reasons for program failure.

The low take-up seems surprising at first sight, as many individuals display considerable work capacity, according to administrative and survey data (see Table 5.A.2, which also contains descriptive statistics for all variables used). For example, 30% of individuals report good or very good health, and 18% report no difficulty in finding employment. Moreover, 52% of individuals suffer from mental diseases, which might only temporarily impair health, at least for some individuals. Predicting the effect of financial incentives, however, requires further steps. Section 5.4 therefore presents a model for the financial incentives, and a simulation of program effects.

[75] Case workers recorded these outcome data specifically for the purpose of this study. No data on the contacts with control group members are available.

5.4 A stylized model and predicted effects of seed capital

5.4.1 A stylized model for the effect of seed capital

We illustrate the basic economic forces at work in a simple static model where individuals maximize utility over consumption (c) and leisure (l). We assume that the relative value of "leisure" increases in an individual's health impairment, but we do not explicitly model the utility function. To create a tractable model, we introduce two short-cuts: First, the model assumes a single level of pension benefits and thus a single notch point. Hence, the model simplifies the Swiss scheme, which contains multiple notch points. Second, we assume that individuals are able to work, and that they are able to perfectly mimic their preferred level of work capacity by choosing their number of work hours. This assumption creates a direct mapping from work hours into disability benefits: Individuals receive disability insurance benefits (b) if hours of work ($L = T - l$, where T denotes the maximum time available for either leisure activities or work) fall below a certain threshold (τ). DI pensioners receive seed capital (s) if they expand work beyond the threshold and thus lose DI benefits.

Our model is static and compares a situation without seed capital ($s = 0$) to a situation with seed capital ($s > 0$). In the absence of seed capital, we expect three types of DI pensioners: The first two types choose boundary solutions, that is, they either choose not to work at all (type 1) or to work exactly at the "cash-cliff" that determines the next lower benefit level (type 2). While individuals choosing the former may have either very high disutility of work or low wages (both may reflect the consequences of a disability) individuals choosing the latter would work more if they did not lose disability benefits. The remaining individuals choose employment at the interior solution with the optimal level of hours of work to the left of the cash-cliff (type 3).

In the seed capital scenario, DI pensioners receive a lump-sum payment if they increase hours of work and lose DI benefits. Two different situations can occur

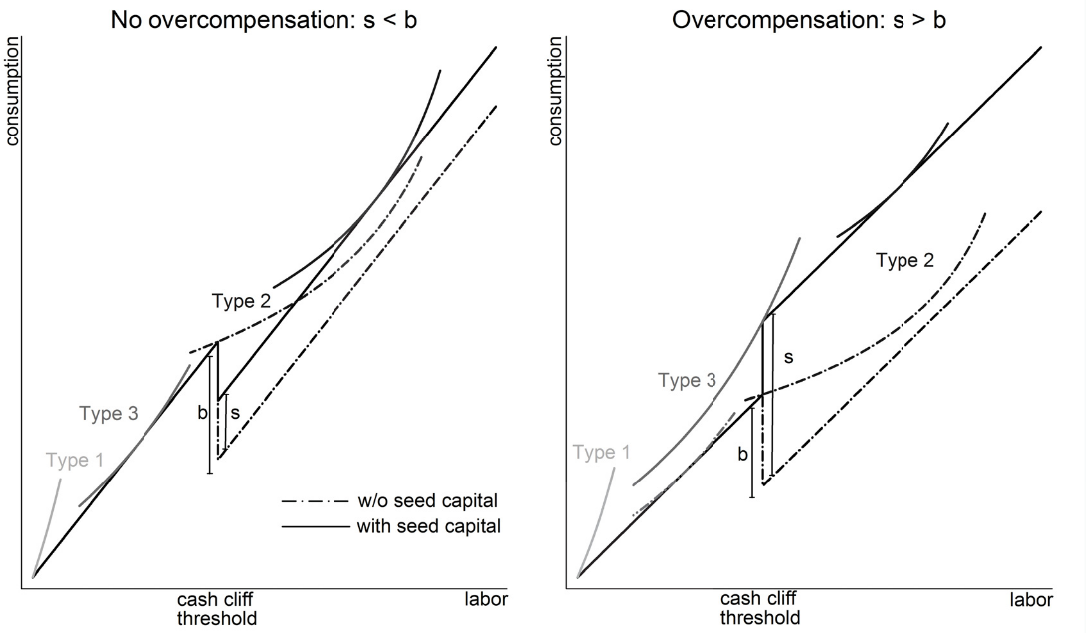
(Figure 5.1): (1) Seed capital does not fully (or just) compensate for the benefit loss (left panel), or (2) seed capital over-compensates for the benefit loss (right panel). In the first case, only individuals who would have chosen their hours of work exactly at the notch point in the absence of seed capital (type 2) change their behavior. They, however, only change their behavior if additional earnings and seed capital together compensate for the loss in benefits and for the higher disutility caused by employment. In other words, total income (earnings, seed capital, and DI benefits) after expanding employment must be strictly higher than total income in the status quo. For all others, the optimal decision remains unchanged (compared to a situation without seed capital). In the second case, that is, if the seed capital over-compensates for the benefit loss, also individuals who choose hours of work below the benefit notch in a world without seed capital react to seed capital. These individuals, however, increase working hours only to the next notch point so that they "just" meet the condition for receiving the seed capital.

The simple model also demonstrates the limits of financial incentives: In the first case discussed above, seed capital increases employment and reduces DI benefits for people of type 1 and type 3 if they are over-compensated for the benefit loss. This implies that the savings in DI benefits due to the intervention are less than the seed capital payments, which cannot be a cost-effective intervention from the perspective of the insurance system. This finding is particularly relevant in the Swiss setting, where individuals receive DI benefits from several sources, while seed capital is paid from the first pillar only. Over-compensation implies that the public pension system (first pillar) "subsidizes" the private occupational pension system (second pillar). Seed capital should thus provide an incentive to expand employment for individuals who are cash-cliff constrained, but should not over-compensate forgone benefits from other sources.

5.4.2 Simulating the financial implication of seed capital

To establish a benchmark for the experimental results, this section presents microsimulation results on the predicted effect of the seed capital offer, based on self-reported income data from the baseline survey.

Figure 5.1: Labor-consumption trade-off



The two panels show labor supply choices in a stylized model under two conditions, for three types of individuals. Left panel: Seed capital does not compensate benefit losses. Right panel: Seed capital over-compensates benefit losses. Notation: s: amount of seed capital, b: loss of benefits if an individual extends his/her earnings beyond a certain cash-cliff threshold. Wages are denoted by w. The budget constraint is $C = wL + b$ if work hours L are below the cash-cliff threshold ($L \leq \tau$) and $C = wL + s$ if individuals expand their work hours a beyond the cash-cliff threshold and claim seed capital.

We model three different "return-to-work" scenarios (henceforth, we use "return-to-work" as a collective term for both "extension of working hours" and "take-up of work"): First, return-to-work for two years; second, return-to-work for five years, and third, return-to-work until retirement. Individuals had the legal possibility to return to their old DI contracts when they were unable to work for 30 consecutive days within the first five years after reintegration (see Section 5.2.2). The first scenario therefore assumes a return-to-work period of two years, where individuals fall back to their old DI contracts after they received the last payment tranche. Yet, a lively political debate on future reforms of the Swiss Disability Insurance Act took place at the time of the experiment, particularly on how to enforce the reintegration of current DI pensioners. DI recipients may thus have feared an abolishment of the fall-back rule. Therefore, the second scenario assumes that individuals increase their employment for a period of five years and fall back to their old disability degree afterwards (but not into their old DI contract, see further explanations below); the third scenario assumes that individuals increase employment until retirement and do not fall back to their old DI degree.

We assume that individuals increase employment exactly to the next cash-cliff threshold. Our data contain current earnings and the disability degree for all working individuals. Cash-cliff thresholds, however, are a function of unobserved potential earnings (see Section 5.2). To construct cash-cliff thresholds, we assume that an individual's current employment level corresponds exactly to his/her disability degree. In other words, if a person had an initial disability degree of 50% and takes up seed capital, his/her employment level increases to 60% and his/her disability degree declines to 40%. This implies that his/her current earnings increase by 20%. For individuals who are currently not working, we predict earnings when taking up employment based on information for individuals who are comparable in terms of observable characteristics, but who are working (see Section 5.C).

During the return-to-work period, increased earnings lead to a reduction in first and second pillar benefits by one quarter. We also recalculate means-tested benefits, as these depend on earnings and on first and second pillar benefits. We assume that those individuals who return to work for two years fall back into their old DI contract. Compared to the status-quo, their DI benefits decline during the return-to-work period, but afterwards, their benefits pick up the status-quo path again. This is not the case when the return-to-work period is five years and longer. Here, the DI recalculates benefits even if individuals fall back into their old disability degree. Furthermore, return-to-work has implications for old-age pensions, which also require recalculation. We provide a detailed description of the simulation in Section 5.B.

Based on the micro-simulation, we estimate necessary return-to-work conditions for different types of individuals. We cannot directly observe types, owing to unobserved work capacity; therefore, we construct types as follows: Type 1 are individuals who do not work at all, irrespective of their disability degree (65% of our sample); type 2 are cash-cliff constrained individuals, that is, individuals who work and have a disability degree exactly at the threshold (12% of our sample); and type 3 are individuals who work and have a disability degree not at the threshold (23% of our sample).

Table 5.1 presents the simulation results. If people perceived that they can fall back to their old DI contract after two years, 14% of the total population would react to low seed capital (CHF 9,000), and almost half of the population would respond to high seed capital (49%). In the low seed capital condition, particularly those who are cash-cliff constrained would react to seed capital (75%). Yet, seed capital rarely over-compensates individuals with longer return-to-work periods. Individuals who are not cash-cliff-constraint would not respond to seed capital. The share of cash-cliff constrained individuals who would take-up seed capital, however, is remarkably stable at around 50%, even in the long run. We thus expect overall small interest in the program if people fear that they cannot return to their old DI contract after two years. However, the interest should be considerably higher among individuals with disability degrees close to threshold values, as these might be cash-cliff constrained.

Table 5.1: Necessary return-to-work condition for alternative scenarios

	Type 1	Type 3	Type 2	Total
Labor market status	*Not working*	*Working*	*Working*	
Disability degree	*Any*	*Not at the notch*	*At the notch*	
% of population	*65%*	*23%*	*12%*	
	Seed capital > benefit loss during return-to-work		Seed capital > total income change	
Percentage where return-to-work condition is fulfilled (9,000/18,000 CHF)				
RTW for 2 years	*7%/41%*	*11%/58%*	*61%/75%*	*14%/49%*
RTW for 5 years	*0%/5%*	*2%/7%*	*53%/58%*	*7%/12%*
RTW until retirement	*0%/2%*	*2%/2%*	*47%/51%*	*6%/8%*

The simulation is based on information from 2,273 individuals in the treatment and control group who participated in the survey and have non-missing information on wages and benefit payments. Individuals who have never worked before DI entry were excluded because wage predictions are based on work history prior DI entry. RTW: Return-to-work. RTW also includes individuals who are already working, but extend their work hours. Details on the simulation can be found in Section 5.B. Source: Own calculations based on administrative and survey data, provided by the Federal Social Insurance Office Switzerland for the purpose of this study.

5.5 Results

5.5.1 Main results

By the end of the field phase, only 0.5% of treated individuals took up seed capital. This result is consistent with our simulation results if individuals perceive that they have to return to the labor market for more than two years. Indeed, we can interpret this fraction as a zero treatment effect: The take-up rate corresponds approximately to the standard rate of pension reduction in previous years. In 2011, about 0.4% of all pensioners reduced their pension payment in comparison to the previous year by at least one quarter, but kept a pension of at least one quarter. Only some of these individuals took up a job or increased employment; most of these pension reductions were not driven by higher incomes. Thus, seed capital might generate wind-fall profits for few people who would have reduced their DI pension anyways, but does not seem to incentivize take-up or expansion of employment.

Short-term reactions further document the low interest in the program. Only 4% of individuals contacted their case worker by phone within the first five months after program announcement to receive more information on the program (see Table 5.A.2). Furthermore, as Table 5.2 reveals, doubling the size of the incentives has no detectable effect either.

We also find little evidence of heterogeneous treatment effects with respect to characteristics that mirror work capacity, such as health, perceived difficulty to find employment, and education (see Table 5.A.3). Only individuals who report that they could easily find a job are slightly more likely to react within the first five months than individuals who report difficulties in job search, but the estimate suffers from a large standard error. Furthermore, the effect of different incentive amounts does not vary strongly with individual characteristics. Only individuals with a college degree are significantly more sensitive to the size of the incentives than individuals without a college degree (significant at the 10%-level), which could point to the role of bounded rationality.[76]

[76] Several studies in behavioral economics show that agents who are faced with complex decisions tend to avoid making an active choice in order not to incur large up-front problem-solving

Table 5.2: Short-term interest in seed capital

	Any contact	Contact and expressed interest	Contact and made appointment
High seed capital	-0.002	-0.002	-0.005
	-0.012	-0.009	-0.008
Constant	0.073***	0.037***	0.033***
	-0.008	-0.006	-0.006
R2	0	0	0
N	4,000	4,000	4,000

The table shows regression results for the 4,000 indviduals who received a seed capital offer. Panels (1)-(3) contain different binary dependent variables. (1): Individual contacted his/her case worker with positive or negative feedback on the letter; (2) Individual asked for information about the program; (3) Individual made an appointment to discuss next steps. The table presents coefficients from OLS regressions with sampling weights. High seed capital: Indicator variable for the high treatment condition (see Section 5.2). The reference category is low seed capital. Standard errors are shown in parentheses. Significance levels: * p<0.1; ** p<0.05; *** p<0.01. Source: Own calculations based on case worker records, provided by the Federal Social Insurance Office, Switzerland.

5.5.2 Bunching behavior and threshold effects

The predictions in Section 5.4 encourage further investigation into the role of cash-cliff constraints and their impact on interest in seed-capital, even if final take-up is low. Our prediction is twofold: First, the stepwise benefit structure should induce bunching at the thresholds, as long as preferences for work versus leisure are smoothly distributed in the population. Second, cash-cliff constrained individuals should react more frequently to the offer.

Figure 5.2 provides empirical evidence in support of these hypotheses. The upper panel displays strong bunching behavior prior to the implementation of seed capital:

costs (Samuelson and Zeckhauser, 1988, Frank and Lamiraud, 2009). Beshears et al. (2008) argue that choices with consequences far in the future are especially complex. Taking up seed capital certainly falls into that category: Determining the consequences of return-to-work on lifetime income requires projecting health, wage and job uncertainty, benefits from different social insurance programs, and capital market returns. It is thus very likely that many DI recipients do not fully understand the lifetime implications of the return to work decision und therefore avoid taking active steps.

An unusually high share of individuals has disability degrees close to a threshold (particularly 50% and 70%), and a low share of individuals has disability degrees just below these thresholds (i.e., 49% and 69%). This pattern might additionally results from rule-of-thumb evaluations of DI degrees by case workers who tend to award DI degrees at prominent numbers (see Section 5.6). The lower panel presents behavioral responses to the announcement of seed capital: Interest in seed capital is typically higher for individuals just above the threshold, compared to individuals just below the threshold. For example, interest in seed capital significantly increases by 0.036 (SD 0.016) at the 50% threshold. The jumps at the other thresholds are smaller and statistically insignificant. These effects are far lower than our simulation predicts.[77]

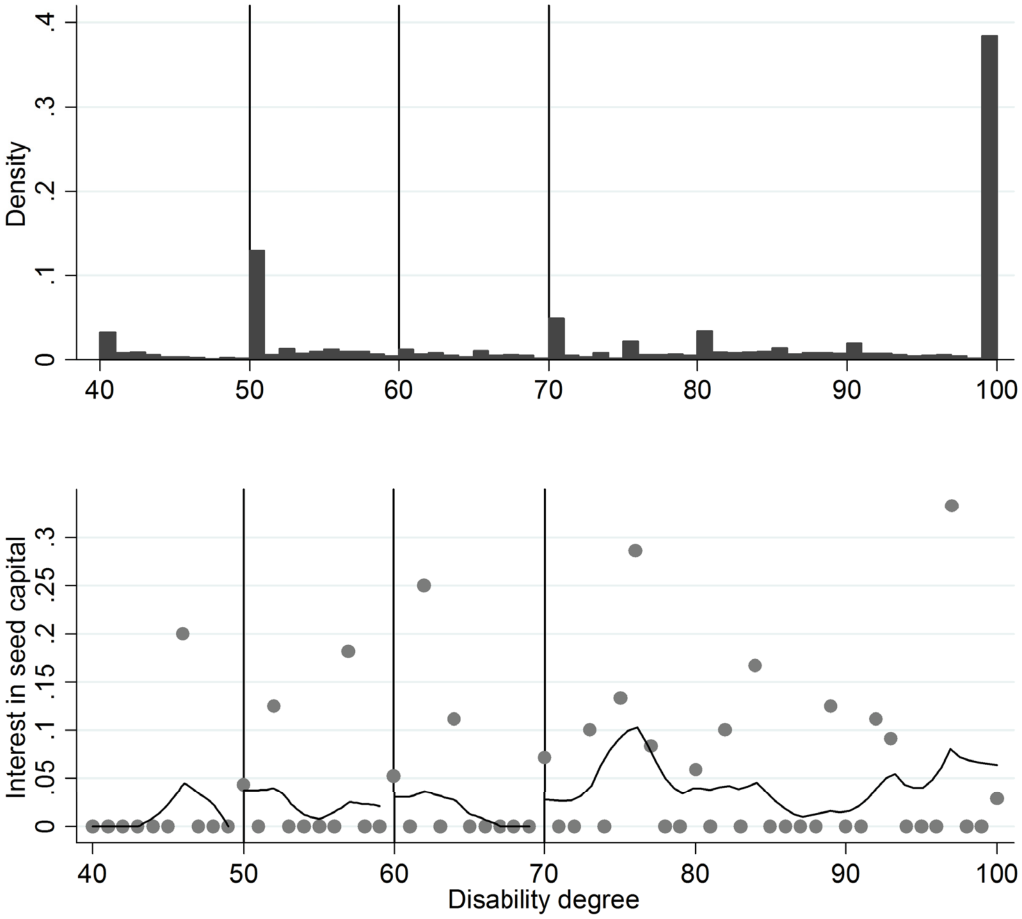
5.6 Discussion: What are the reasons for low interest and take-up?

This section provides three potential reasons for the low take-up. First, risk-aversion might explain the low interest in seed capital. To take up seed capital, individuals need to trade off a relatively safe DI insurance payment against a potentially higher, but more uncertain, work income. Risk aversion could thus significantly harm the expansion of employment and the take-up of seed capital, particularly for individuals with longer return-to-work periods (see Section 5.4.2).

Second, our interpretation of bunching behaviour as response to a non-linear budget set might be flawed. Since we observe disability degrees and labor supply decisions, but not individual earnings thresholds, we cannot conclude with certainty that individuals with disability degrees at thresholds really bunch at their individual earnings thresholds. Individuals commonly bunch at disability degrees that are not associated with higher DI benefits (for example, 80% and 100%). Thus, clustering of disability degrees at decimal numbers could reflect rules of thumb that guide case

[77] Interest in low and high seed capital is combined due to sample size restrictions. The estimates for the other two notch points are 0.029 (SD 0.032) for the 60% threshold and 0.032 (SD.0243) for the 70% threshold.

Figure 5.2: Bunching behavior and responses to seed capital at the cash-cliffs



The figure is based on information from respondents who participated in a survey prior to the pilot project, who were employed prior to the experiment, who provided survey information on earnings, and who were randomized into one of the treatment groups (N = 760). The upper panel presents a histogram of disability degrees with bin width of one percentage point. The lower panel presents interest in seed capital (binary variable: individual contacts local disability office and expresses interest). Dots are averages per disability degree; lines represent the results of kernel-weighted local regression using a triangle kernel and a bandwidth of 3. Source: Own calculations based on administrative and survey data and case worker records, provided by the Federal Social Insurance Office, Switzerland.

workers' assessment of the disability degree, rather than true labour supply effects. Consequently, the true proportion of individuals who are cash-cliff constrained could be much smaller than expected. If individuals at notch points were not cash-cliff constrained, we would predict much lower take-up rates.

Third, low take-up could reflect information frictions (Bhargava and Manoli, 2013, Currie, 2006). A letter announced the program to DI recipients (see Section 5.D), but we cannot track whether individuals opened, read, and processed the letter, as reacting to the letter was completely voluntary. Yet, from case worker records, we know that 8% of individuals contacted their DI caseworkers. Half of these individuals, or 4% in total, called their case worker to reject the offer. 1% of individuals asked the caseworker to explain the content of the letter, and 3% showed interest in the program. Unfortunately, we are not able to track individuals who did not reply at all.

5.7 Conclusion

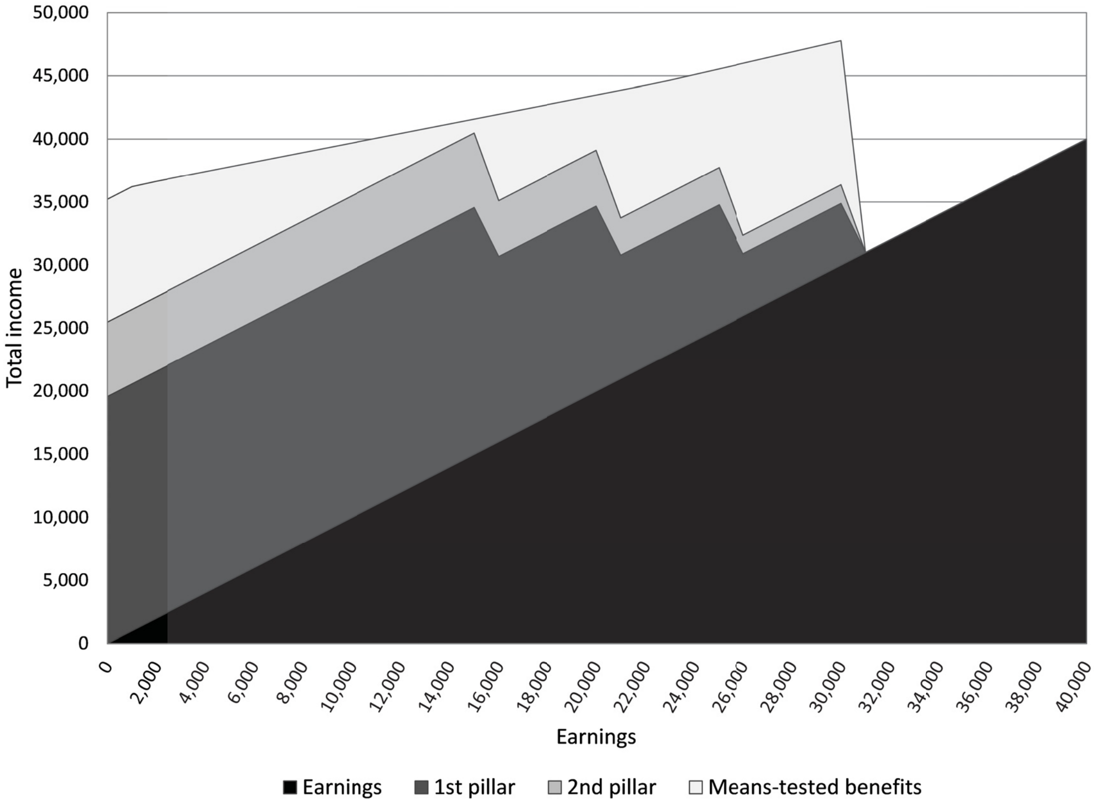
This paper presents the results of a field experiment on financial work incentives for DI recipients in Switzerland. The program aimed at reducing the loss of DI benefits if earnings exceed certain thresholds ("cash-cliffs"). The program granted a substantial lump-sum payment of up to CHF 72,000 (USD 71,000) if individuals expanded employment and thus reduced their DI payments.

Overall take-up of the financial incentives was low at 0.5%. Furthermore, only few individuals (4%) contacted their case worker within five months after the program announcement to learn more about the program. These low numbers are consistent with findings from a micro-simulation: For a large majority of individuals, returning to the labour market for a period of more than two years would not have been beneficial. We examine the reactions to the program by subgroups, using case-worker records on interactions with treated individuals within the first five months: First, individuals with better subjective health status or with better different employment opportunities are not more likely to react to the program announcement. Second, doubling the amount of the incentives made no difference either. Third, individuals who desire to work more, but who face particularly strong work disincentives ("cash-cliff constrained" individuals) are more likely to react. These effects, however, are much smaller than corresponding micro-simulation results would suggest. We thus conclude that the share of individuals that are truly cash-cliff constrained is much smaller than we initially expected; instead, bunching behaviour at threshold values might result from rule-of-thumb behaviour of caseworkers when they classify an individuals' work capacity. Moreover, risk-aversion, bounded rationality, and information frictions might have reinforced the low interest in the conditional cash transfer program.

Appendix

5.A Figures and tables

Figure 5.A.1: Budget constraint for a single DI recipient



The figure shows the predicted household income of an example household (disability benefit recipient in a single household), depending on his/her earnings on the first labor market. Assumption: Potential earnings amount to 50,000 CHF if the individual worked fulltime. The x-axis states earnings on the first labor market in CHF per year. The y-axis states the total income, including 1st pillar benefits, 2nd pillar benefits, and means-tested benefits in CHF per year (see Section 2 for definitions of these income sources). Source: Bütler et al. (2012), p. 186.

Table 5.A.1: Sampling structure

	Obs.	% full sample	Stratified
1) Full sample	37,853	100%	No
2) Invited for survey participation	8,000	21%	Yes
3) Survey participants	4,049	11%	Yes
Nonparticipants	3,951	10%	Yes
4) Experimental sample	6,020	16%	Yes
Seed capital high	2,000	5%	Yes
Seed capital low	2,000	5%	Yes
Control group	2,020	5%	Yes
5) Simulation sample	2,273	6%	Yes

Selection for participation took place in two steps: From the total of 37,853 individuals who were observed in the administrative records in June 2009, 2,814 individuals have been excluded, primarily as their current residence was outside of the cantons of St. Gallen and Vaud. From the remaining 35,039 individuals, 8,000 individuals have been randomly selected to participate in a survey. Random sampling was stratified by three age groups. The experimental sample consists of all individuals who were invited to participate in the survey, but excluded individuals who are likely to live in a nursing home, and individuals with a disabled partner (to avoid spill-over effects if one person gets randomized into the low and the other person gets randomized into the high seed capital group). The simulation sample consists of all individuals in the treatment and comparison group who participated in the survey and have non-missing information on incomes and benefit payments. Individuals who have never worked before DI entry were excluded, because wage predictions are based on work history prior to disability.

Table 5.A.2: Descriptive statistics

	Obs.	Mean
Phone call: Positive/neutral reaction[1]	4,000	0.04
Phone call: Any reaction[1]	4,000	0.08
Phone call: Only positive reaction[1]	4,000	0.03
Seed capital: low[1]	4,000	0.5
Seed capital: high[1]	4,000	0.5
Type 1: not working[2]	2,297	0.63
Type 2: working at notch[2]	2,297	0.1
Type 3: working not at notch[2]	2,297	0.27
Total yearly benefit level (in 1,000 CHF)[2]	1,813	31.77
Yearly wage (in 1,000 CHF)[2]	2,202	6.24
Self-reported health: good/very good[3]	2,198	0.31
Has any pains[3]	2,200	0.77
Difficulty: Mobility[3]	2,206	0.4
Difficulty: Household[3]	2,214	0.6
Difficulty: Self-care[3]	2,214	0.2
Years in DI[3]	2,214	0.06
No difficulty to find new employment[3]	2,214	0.18
Age[3]	2,214	42.19
Male[3]	2,214	0.48
Foreign[3]	2,214	0.31
Civil status: Single/widow[3]	2,214	0.43
Civil status: Married[3]	2,214	0.41
Civil status: Divorced/separated[3]	2,214	0.16
Dependent children[3]	2,214	0.37
Disease: Mental[3]	2,214	0.52
Disease: Nervous system[3]	2,214	0.08
Disease: Back disorders[3]	2,214	0.06
Disease: Other musculoskeletal diseases[3]	2,214	0.09
Disease: Injuries[3]	2,214	0.09
Disease: Other[3]	2,214	0.16
Start of pension receipt: Before 1996[3]	2,214	0.22
Start of pension receipt: 1996 - 2000[3]	2,214	0.25
Start of pension receipt: 2001 - 2006[3]	2,214	0.36
Start of pension receipt: After 2006[3]	2,214	0.18

Continued on the next page.

Continued from the previous page: Descriptive statistics

	Obs.	Mean
Education: Compulsory education or less^3	2,214	0.35
Education: Vocational degree^3	2,214	0.52
Education: High school degree^3	2,214	0.04
Education: Higher vocational or college^3	2,214	0.09

The table presents descriptive statistics for the sample of treated individuals, or for subgroups with non-missing information on the respective variables. Samples: (1) Individuals in both treatment groups; (2) Individuals in treatment groups with survey response; (3) Individuals in sample 2 with non-missing information on capacity-to-work variables (such as difficulty to find employment). Source: Own calculations based on administrative data and survey data, provided by the Federal Social Insurance Office, Switzerland.

Table 5.A.3: Effect heterogeneity

	# Obs.	Intercept	High seed capital
Self-rated health			
good/very good	708	0.049** (0.015)	0.001 (0.024)
fair/bad	1,569	0.042*** (0.011)	-0.01 (0.015)
P-value (difference)		0.749	0.69
Difficulty to find employment			
Easy	138	0.086 (0.064)	-0.052 (0.066)
Difficult	2,159	0.042*** (0.009)	-0.001 (0.013)
P-value (difference)		0.487	0.454
Education			
Higher education	210	0.026 (0.013)	0.065 (0.042)
No higher education	2,087	0.047*** (0.010)	-0.014 (0.013)
P-value (difference)		0.212	0.071

The regression coefficients are from an OLS regressions with survey weights using information from treatment groups. Reference category is low seed capital. Standard errors are shown in parentheses. Significance levels: * p<0.1; ** p<0.05; *** p<0.01* p<0.1; ** p<0.05; *** p<0.01

5.B Simulation

This appendix describes the assumptions and procedures used to simulate the return-to-work incentives described in the main text. Our sample for this analysis consists of all individuals in the treatment or the comparison groups who participated in the survey and have non-missing information on other sources of income (i.e. means-tested benefits, second pillar benefits, and spousal earnings). We also exclude recipients who have not been employed prior to DI entry, because we rely on the employment history prior to disability to predict earnings in case a DI recipient returns to work. With these restrictions, we have a final sample of 2,273 DI recipients (see Table 5.A.1 in the appendix).

Return-to-work incentives are measured by comparing the net present discounted value of lifetime income under the status-quo with a situation in which DI recipients reduce their disability benefits by a quarter of a full disability pension and take up or expand employment. The difference in lifetime income is calculated as follows:

$$\begin{aligned}\Delta\text{income} = \sum_{t=0}^{T-a_0} \pi_{t|0} * \left(\frac{1}{1+r}\right)^t * [d * (w_t^{dur} + b_t^{dur} + p_t^{dur} + m_t^{dur}) + (1-d) \\ *(w_t^{post} + b_t^{post} + p_t^{post} + m_t^{post}) - w_t^{quo} - b_t^{quo} - p_t^{quo} - m_t^{quo}] \quad (5.1)\end{aligned}$$

where a_0 is the age today, π is the probability for being alive at some future date t conditional on being alive today, r is the interest rate, and d is a dummy which is 1 during the return-to-work period and 0 otherwise.[78] The variables w_t^{quo}, b_t^{quo}, p_t^{quo}, and m_t^{quo} measure earnings, first pillar benefits, second pillar benefits, and means-tested benefits in period t under the status quo. Similarly, the variables w_t^{dur}, b_t^{dur}, p_t^{dur}, m_t^{dur} and w_t^{post}, b_t^{post}, p_t^{post}, m_t^{post} measure earnings, first pillar benefits, second pillar benefits, and means-tested benefits during return-to-work ($d = 1$) and after return-to-work ($d = 0$), respectively.

[78] We assume a real interest rate of 2.5% and a maximum life span T of 100 years. Survival probabilities are taken from the age and sex specific life tables published by the Swiss Federal Statistical Office (http://www.bfs.admin.ch/bfs/portal/en/index/themen/01/02/blank/dos/la_mortalite_en_suisse/tabl01.html).

Earnings

Earnings of DI recipients under the status quo, w_t^{quo}, can be observed directly in the data. We assume that DI recipients continue to work at the same level until they reach the full retirement age when they permanently leave the labor force. Earnings adjust over time with the growth rate g=1%, which corresponds roughly to the real wage growth rate in Switzerland during the past 20 years. Computing the earnings during the return-to-work period w_t^{dur} requires projecting the DI recipient's potential earnings when rejoining the workforce. We use the earnings information from DI recipients who are currently working to estimate potential earnings for all DI recipients using a regression-based imputation procedure (see Appendix 5.C for a detailed description). We assume that during the return-to-work period DI recipients work the maximum percent they are allowed to work before their benefits get cut. For example, a DI recipient who during the return-to-work period receives a quarter of a full disability pension works 60 percent of a full time job. Finally, earnings in each year after return-to-work w_t^{post} are assumed to be equal to the earnings under the status quo in that year.

First pillar benefits

First pillar benefits under the status quo b_t^{quo} can be observed directly in the administrative records and adjust over time based with the earnings growth rate g.[79] During the return-to-work period first pillar benefits are reduced by one quarter of a full DI pension $b_t^{dur} = b_t^{quo}/x_t^{quo} * x_t^{dur}$, where x_t^i denotes the fraction of a full disability pension that a beneficiary receives in year t (x_t^i = 0, 0.25, 0.5, 0.75, 1) and $x_t^{dur} = x_t^{quo} - 0.25$.

In the case in which recipients return-to-work for two years disability benefits after return-to-work b_t^{post} are equal to b_t^{quo}. If the return-to-work period lasts five years or more, disability benefits after return-to-work are re-calculated taking into

[79] According to the law, wage growth and inflation have an equal weight in the indexation of first pillar pensions and means-tested benefits. Because the wage growth rate was approximately equal to the inflation rate in the past decades, ignoring the inflation rate in the indexation formula is not crucial.

account the earnings and contributions during the return-to-work period. More specifically, b_t^{post} is calculated using the piecewise linear formula

$$b_t^{post} = x_t^{post} * f(q_t^{post}) * \begin{cases} \underline{b} & \text{if } v_t^{post} \leq \underline{b} \\ 0.74 * \underline{b} + \frac{13 * v_t^{post}}{600} & \text{if } \underline{b} < v_t^{post} < 3 * \underline{b} \\ 1.04 * \underline{b} + \frac{8 * v_t^{post}}{600} & \text{if } 3 * \underline{b} \leq v_t^{post} \leq 6 * \underline{b} \\ 2 * \underline{b} & \text{if } v_t^{post} > 6 * \underline{b}, \end{cases}$$

where $\underline{b}$ is the minimum pension, v_t^{post} is the assessment basis, and $f(q_t^{post})$ is an adjustment factor, which is increasing in the number of contribution years q_t^{post}. The assessment basis is determined by the average earnings in all years (uncapped) after applying revaluation factors to adjust for wage inflation. Prior to the statutory retirement age x_t^{post} is equal to x_t^{dur}. After the statutory retirement age DI recipients qualify for a full pension, so that x_t^{post} is equal to 1.

Second pillar benefits

Around 39% of DI beneficiaries in our sample receive DI benefits from the occupational pension scheme (second pillar). Second pillar DI benefits under the status quo p_t^{quo} can be observed in the data and are assumed to adjust over time with the earnings growth rate g. During the return-to-work period the second pillar DI pension is reduced by one quarter of a full second pillar DI pension $p_t^{dur} = p_t^{quo}/x_t^{quo} * x_t^{dur}$, where $x_t^{dur} = x_t^{quo} - 0.25$.

As for the first pillar, second pillar benefits in the after return-to-work period p_t^{post} are equal to p_t^{quo} if recipients return-to-work for less than five years. If the return-to-work period exceeds five years, p_t^{post} is re-calculated using the following formula:

$$p_t^{post} = p_t^{dur} + (x_t^{post} - x_t^{dur}) * cr * k_t^{post}, \tag{5.2}$$

where cr is the conversion rate (equal to 7%) at which accumulated capital k_t^{post} during the return-to-work period is translated into a lifelong pension. The accumulated capital k_t^{post} consists of all contributions made during the return-to-work period plus hypothetical contributions that the individual would have made until the statutory retirement age if his health status had not deteriorated. Because recipients only receive the fraction of a full disability pension that they have forgone during the return-to-work period in addition to p_t^{dur}, the full second pillar disability pension based on the contributions during the return-to-work period ($cr * k_t^{post}$) is adjusted by the factor ($x_t^{post} - x_t^{dur}$). After the statutory retirement age recipients receive a full disability pension, which is equal to $p_t^{dur} + cr * k_t^{post}$.

Means tested benefits

In our sample, around 32% of DI beneficiaries claim means-tested benefits, which are awarded in case DI benefits from the first and second pillar are not sufficient to meet minimial costs of living. Means-tested benefits under the status quo m_t^{quo} can be observed directly in the data and adjust over time with the earnings growth rate g. The calculation of means-tested benefits during and after the return-to-work period requires knowledge of a recipient's income, assets as well as total expenditures (cost-of-living allowance, rent or interest on mortgage, and health care). We observe a recipient's income and cost-of-living allowance, but we have no information on assets, rent or mortgage payments, and health care expenditures that are not covered by the mandatory health insurance.

To surmount this problem, we use the following approach: First, we calculate the hypothetical annual means-tested benefits $\hat{m}_t^{quo}$ ignoring potential asset holdings and health care expenditures that are not covered by the health insurance:

$$\hat{m}_t^{quo} = \max\left(l_t + h_t + s_t - b_t^{quo} - p_t^{quo} - 0.66 * e_t - \max(0.66 * w_t^{quo} - z_t; n_t); 0\right), \tag{5.3}$$

where l_t is a cost-of-living allowance, h_t denotes the health insurance premium, s_t denotes expenditure for housing, and e_t denotes spousal earnings. The calculation of means-tested benefit also includes hypothetical earnings n_t or two thirds of a DI recipient's earnings w_t^{quo} less an exemption z_t whichever is higher. The level of hypothetical earnings n_t depends on a DI recipient's remaining work capacity.

Second, we calculate an adjustment factor adj_t by subtracting the actual annual means-tested benefits in the status quo m_t^{quo} from the hypothetical annual means-tested benefits $\hat{m}_t^{quo}$:

$$adj_t = \hat{m}_t^{quo} - m_t^{quo} \tag{5.4}$$

The adjustment factor thus measures the bias in the amount of hypothetical means-tested benefits that is due to asset holdings and health care expenditures. Third, if we assume that asset holdings and health expenditures are unaffected by the return-to-work decision, then we can calculate means-tested benefits during and after return-to-work according to the following formula:

$$m_t^i = \hat{m}_t^i - adj_t \text{ for } i = dur, post. \tag{5.5}$$

5.C Imputation of earnings

Potential earnings when taking up seed capital (w_t^{dur}) are unobserved. To predict earnings for all DI recipients, we implement a regression-based imputation procedure based on earnings information from DI recipients who are currently working. We proceed in three steps:

Step 1: Predicting potential earnings

The disability degree determines the percentage loss in earnings due to disability i.e. is computed by the DI office as

$$\text{DI degree} = 1 - \frac{\text{Potential earnings w/disability}}{\text{Potential earnings w/o disability}}. \tag{5.6}$$

Rewriting equation 5.6 gives the hypothetical income of an individual if the individual was not disabled:

$$\text{Potential earnings w/o disability} = \frac{\text{Potential earnings w/disability}}{1 - \text{DI degree}}. \tag{5.7}$$

We assume that individuals can fully mimic their disability degree by signaling their potential earnings with disability. Then, potential earnings without disability equal their current earnings divided by 1 minus the disability degree:

$$\text{Potential earnings w/o disability} = \frac{\text{Current earnings}}{1 - \text{DI degree}}. \tag{5.8}$$

If individuals take up seed capital, the disability degree has to decrease, and current earnings must increase accordingly (potential earnings w/o disability are assumed to remain constant over time). Denote the new level of current earnings in case of seed capital take-up as "Current earnings$_{sc}$", and the new disability degree as "DI degree$_{sc}$".

Rewriting equation 5.8 gives an expression for "Current earnings$_{sc}$" under seed capital take-up:

$$\text{Current earnings}_{sc} = \text{Potential earnings w/o disability} * (1 - \text{DI degree}_{sc}). \quad (5.9)$$

Computation of "Current earnings$_{sc}$" would be straightforward for individuals who are currently working: We can compute potential earnings without disability from equation 5.8 and plug them into equation 5.9.[80] We can then compute "Current earnings$_{sc}$" for different levels of "DI degree$_{sc}$".

Yet, for individuals who are not working prior to the experiment, current earnings are zero, but potential earnings without disability are not. We therefore impute potential earnings without disability for the full simulation sample. We start by estimating the following model for the sample of DI recipients who are currently working.

$$\ln(\text{potential earnings w/o disability}_i) = \alpha + \beta X_i + \epsilon_i, \quad (5.10)$$

where potential earnings without disability are computed according to equation 5.8, X_i is a vector of explanatory variables often used to predict earnings such such as gender, nationality, civil status, children, disability, health, pension payment and start of pension, number of years contributed to the pension system before inflow into disability insurance, average labor income before inflow into disability, log workload per week (workload is measured in hours as a fraction of 42 hours), and education. We use all observations from individuals who were employed at the time of the baseline interview, reported their wages, do not work in sheltered workshops (since their wage does not represent market wages), and report plausible hours of work (in total 561 individuals). Results are not reported but available from the authors upon request.

[80] Hence, potential earnings without disability are not defined for individuals with a DI degree of 100%.

Step 2: Predicting workload

The coefficients from the above regression are used to to predict potential earnings without disability. All explanatory variables are observed in the data. However, workload is unobserved (or zero) for those who are not working. Workload must therefore be predicted for those who are not working.

We use the following regression to predict workload:

$$\ln(\text{workload}_i) = \gamma + \delta X_i + \nu_i, \tag{5.11}$$

where X_i is a vector of explanatory variables that is identical to the vector of variables used in equation 5.10, except for ln(workload), which is now the dependent variable. The results are not reported but available from the authors upon request.

Step 3: Imputing potential earnings without disability

In order to impute potential earnings without disability, we compute fitted values from regression 5.10 for all individuals in the sample. For individuals who are currently working, all regressors are taken from administrative and survey data, including workload. For those individuals who are not working, we plug in the fitted values obtained in Step 2 for workload to replace missing values (or zeroes) for workload.

In order to capture the uncertainty associated with the computation of fitted values for potential earnings without disability, we compute a distribution of potential wages without disability for each individual. More specifically, for each individual, we randomly draw 1,000 error terms derived from regression 5.10 and add them to their fitted values in order to obtain 1,000 values for potential earnings without disability.

5.D Offer letter

This section contains the English translation of the seed capital offer letter. The original languages were German and French. For the original version, please contact the authors.

Dear Mr./Mrs. Miller,
Many disability insurance recipients wish to take up work or to extend their working hours. In many cases, however, starting a job or extending an existing work relationship is associated with financial losses. Therefore, the Swiss disability insurance wants to give some benefit recipients the possibility to receive a seed capital if they start a job and therefore manage to reduce their disability insurance benefit receipt. In this way, the Swiss disability insurance wants to ease the negative financial consequences of employment or extension in working hours.

You belong to the group of people that are selected to participate in the project. If you feel able to take up a job or to extend your working hours, and if your pension decreases as a consequence, you will have the possibility to receive a payment. You will find more information on the amount of the payment and on your eligibility in the attachment.

Participation in the project is voluntary. You will not incur any disadvantages if you cannot or do not want to accept the offer. In this case, please regard this letter as irrelevant. Your current rights and obligations will remain unchanged.

Are you interested in participating in the project, or do you have any questions? Please contact your disability insurance office directly. The office will help you in your efforts.
[Phone number of disability insurance case worker]
Kind regards,
N.N., Director of the disability insurance office

Supplementary information sheet

What is "seed capital"?
In many cases, starting a job or extending an existing work relationship is associated with financial losses for disability insurance recipients. Means-tested benefits ("Ergänzungsleistungen") as well as second pillar benefits ("Leistungen der Beruflichen Vorsorge") might decrease. Therefore, your new income might be smaller than the combined benefits from your pension and these other sources. Disability insurance benefit recipients who participate in the project are eligible for a payment. Two conditions have to be satisfied: First, the recipient has to take up a job in the regular labor market, or extend his job in the regular labor market. Second, as a result of taking up or extending a job, his pension has to be adjusted downwards in the course of an official revision.

Who can participate in the pilot project?
The aim of the project is to evaluate the seed capital program. Therefore, only those people who received this letter are eligible to participate. You belong to this group.

Do I have to participate in the pilot project?
Participation is completely voluntary. If you are not able to participate due to your health status, or if you would like to abstain from participating for other reasons, you do not have to participate. If you decide not to participate in the pilot project, you do not need to do anything. Not participating does not have any disadvantages. Your rights and obligations will be unchanged.

What do I have to do if I would like to participate in the pilot project?
The pilot project lasts until Juli 31st, 2013. If you would like to participate in the pilot project and if you have further questions, please contact your local disability insurance office at [Phone number of disability insurance case worker]. If you would like to participate in the pilot project and you do not have any questions, please report your new employment status until December 31st, 2012, to your disability insurance office. Please include a copy of this letter as well as a copy of your work contract to your report.

Which support am I going to receive if I would like to take up employment or increase my working hours?
Participating in the project implies that you will take the initiative to find a job. Of course, you are eligible for support of your disability insurance office as usual. Please contact your disability insurance office for support and help.

How and when will the seed capital payment be made?
The seed capital will be paid after you take up employment or extend your working hours, and after the disability insurance office has confirmed your pension reduction. You will be eligible for payment, whether you are employed or self-employed. The seed capital will be paid in tranches. In order to receive payment of the first tranche, the employment relationship has to be in place.

Seed capital and regular revision of your pension
Your eligibility for disability insurance benefits is revised regularly (every 3-5 years). If your regular revision falls into the time of the pilot project and if your pension will be reduced during this revision or even cancelled, the following rules apply: A seed capital will always be paid if the above mentioned conditions are satisfied and the working contract has been signed prior to the regular revision.

How large is the seed capital amount?
The seed capital amount depends on the reduction in your pension. If you currently receive a full pension and your pension is reduced to...

- a three-quarter pension, you will receive a seed capital of 9,000 (18,000) Swiss Francs.
- a semi pension, you will receive a seed capital of 18,000 (36,000) Swiss Francs.
- a quarter pension, you will receive a seed capital of 27,000 (54,000) Swiss Francs.
- no pension, you will receive a seed capital of 36,000 (72,000) Swiss Francs.

(Note: The amount stated is the amount for individuals in the low (high) seed capital condition; Individuals are not aware of different treatment conditions. They will only see the amount that they are eligible for.)
If you currently receive a three-quarter pension and your pension is reduced to...

- a semi pension, you will receive a seed capital of 9,000 (18,000) Swiss Francs.
- a quarter pension, you will receive a seed capital of 18,000 (36,000) Swiss Francs.
- no pension, you will receive a seed capital of 27,000 (54,000) Swiss Francs.

If you currently receive a semi pension and your pension is reduced to...

- a quarter pension, you will receive a seed capital of 9,000 (18,000) Swiss Francs.
- no pension, you will receive a seed capital of 18,000 (36,000) Swiss Francs.

If you currently receive a quarter pension and your pension is reduced to...

- no pension, you will receive a seed capital of 9,000 (18,000) Swiss Francs.

The payment is due in four tranches, and each tranche is due after 6 months. The payment depends on whether the reduction in your pension pertains. Regarding the computation of means tested benefits, the seed capital counts as assets and not as income. For more information on the effect of seed capital on means tested benefits, please contact your local disability insurance office.

What happens if my health status decreases again?
If you can prove that your health status has decreased again, you will be eligible for your old pension. This eligibility rule will apply for five years after the decrease in disability benefits. If your pension increases during the receipt of seed capital, no further tranches will be paid. In this case, however, you do not have to pay back the amount that you have already received. For means tested benefits and second pillar benefits, no general rules exist. In this case, please contact your disability insurance office.

What happens if I lose my job?

If you lose your job for reasons other than your health status (e.g. for operational reasons), your eligibility for seed capital will continue. Your pension as well as your second pillar benefits will remain reduced. In this case, you will be treated like someone whose pension has been reduced in the course of a regular revision. Your advantage will be that you will still receive your seed capital after losing your job.

What is the legal basis for seed capital?

The disability insurance is obliged to bring their clients back into work. In order to test potential programs for the future, the insurance can conduct pilot projects: Art. 68quater IVG. There are no rights impairments of the insured due to pilot projects.

Bibliography

Adam, S., A. Bozio, and C. Emmerson (2010): "Reforming disability insurance in the UK: evaluation of the pathways to work programme," Working paper, Institute for Fiscal Studies, London.

Altonji, J. G. and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1102.

Angrist, J. D. and K. Lang (2004): "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program," *American Economic Review*, 94, 1613–1634.

Arcidiacono, P. and S. Nicholson (2005): "Peer effects in medical school," *Journal of Public Economics*, 89, 327–350.

Autor, D. H. and M. G. Duggan (2007): "Distinguishing Income from Substitution Effects in Disability Insurance," *American Economic Review*, 97, 119–124.

Back, M. D., S. C. Schmukle, and B. Egloff (2008): "Becoming Friends by Chance," *Psychological Science*, 19, 439–440.

Baker, M., J. Gruber, and K. Milligan (2008): "Universal Child Care, Maternal Labor Supply, and Family Well-Being," *Journal of Political Economy*, 116, 709–745.

Behncke, S., M. Frölich, and M. Lechner (2009): "Targeting Labour Market Programmes - Results from a Randomized Experiment," *Swiss Journal of Economics and Statistics (SJES)*, 145, 221–268.

BERLINSKI, S. AND S. GALIANI (2007): "The effect of a large expansion of pre-primary school facilities on preschool attendance and maternal employment," *Labour Economics*, 14, 665–680.

BERTRAND, M. (2011): *New Perspectives on Gender*, Elsevier, vol. 4 of *Handbook of Labor Economics*, chap. 17, 1543–1590.

BERTRAND, M., C. GOLDIN, AND L. F. KATZ (2010): "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, 2, 228–255.

BESHEARS, J., J. J. CHOI, D. LAIBSON, AND B. C. MADRIAN (2008): "How are preferences revealed?" *Journal of Public Economics*, 92, 1787–1794.

BETTINGER, E. P. AND B. T. LONG (2009): "Addressing the Needs of Underprepared Students in Higher Education: Does College Remediation Work?" *Journal of Human Resources*, 44, 736–771.

BHARGAVA, S. AND D. MANOLI (2013): "Why are Benefits Left on the Table? Assessing the Role of Information, Complexity, Stigma on Take-Up with an IRS Field Experiment," Mimeo.

BHATTACHARYA, D. (2009): "Inferring Optimal Peer Assignment From Experimental Data," *Journal of the American Statistical Association*, 104, 486–500.

BHATTACHARYA, D. AND P. DUPAS (2012): "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 167, 168–196.

BLACK, S. E. (1999): "Do Better Schools Matter? Parental Valuation Of Elementary Education," *The Quarterly Journal of Economics*, 114, 577–599.

BLAU, D. AND J. CURRIE (2006): *Pre-School, Day Care, and After-School Care: Who's Minding the Kids?*, Elsevier, vol. 2 of *Handbook of the Economics of Education*, chap. 20, 1163–1278.

Blau, D. M. and P. K. Robins (1988): "Child-Care Costs and Family Labor Supply," *The Review of Economics and Statistics*, 70, 374–81.

Blundell, R. and J. L. Powell (2003): "Chapter 8 - Endogeneity in nonparametric and semiparametric regression models," in *Advances in Economics and Econometrics*, ed. by M. Dewatripont and L. P. Hansen, Cambridge University Press, vol. 2 of *Econometric Monograph Series 36*, 312–357.

Bratti, M., E. D. Bono, and D. Vuri (2005): "New Mothers' Labour Force Participation in Italy: The Role of Job Characteristics," *LABOUR*, 19, 79–121.

Bütler, M., L. Inderbitzin, J. Schulz, and S. Staubli (2012): "Die Auswirkungen bedarfsabhängiger Leistungen: Ergänzungsleistungen in der Schweiz," *Perspektiven der Wirtschaftspolitik*, 13, 137–265.

Campolieti, M. and C. Riddell (2012): "Disability policy and the labor market: Evidence from a natural experiment in Canada, 1998-2006," *Journal of Public Economics*, 96, 306–316.

Card, D. and A. B. Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–93.

Card, D. and J. Rothstein (2007): "Racial segregation and the black-white test score gap," *Journal of Public Economics*, 91, 2158–2184.

Carrell, S. E., R. L. Fullerton, and J. E. West (2009): "Does Your Cohort Matter? Measuring Peer Effects in College Achievement," *Journal of Labor Economics*, 27, 439–464.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013): "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, 81, 855–882.

Cascio, E. U. (2009): "Maternal Labor Supply and the Introduction of Kindergartens into American Public Schools," *Journal of Human Resources*, 44, 140–170.

CONNELLY, R. (1992): "The Effect of Child Care Costs on Married Women's Labor Force Participation," *The Review of Economics and Statistics*, 74, 83–90.

CURRARINI, S., M. O. JACKSON, AND P. PIN (2009): "An Economic Model of Friendship: Homophily, Minorities, and Segregation," *Econometrica*, 77, 1003–1045.

CURRIE, J. (2006): "The Take-up of Social Benefits," in *Poverty, The Distribution of Income, and Public Policy*, ed. by A. Auerbach, D. Card, and J. Quigley, New York: Russel-Sage, 80–148.

DE GIORGI, G., M. PELLIZZARI, AND S. REDAELLI (2010): "Identification of Social Interactions through Partially Overlapping Peer Groups," *American Economic Journal: Applied Economics*, 2, 241–275.

DE JONG, P., M. LINDEBOOM, AND B. VAN DER KLAAUW (2011): "Screening Disability Insurance Applications," *Journal of the European Economic Association*, 9, 106–129.

EGGER, S. AND T. DYLLICK (2010): "Studieren an der HSG: Graduate Survey Report 2010," University of St. Gallen, Stelle für Qualitätsentwicklung.

EPPLE, D. AND R. E. ROMANO (2011): "Chapter 20 - Peer Effects in Education: A Survey of the Theory and Evidence," North-Holland, vol. 1 of *Handbook of Social Economics*, 1053–1163.

ERNST, M. D. (2004): "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, 19, 676–685.

FELFE, C. (2012): "The motherhood wage gap: What about job amenities?" *Labour Economics*, 19, 59–67.

FELFE, C., M. LECHNER, R. ITEN, S. SCHWAB, S. STERN, AND P. THIEMANN (2013): "Familienergänzende Kinderbetreuung und Gleichstellung," Bern: Schweizer Nationalfonds.

Fisher, R. A. (1971): *The Design of Experiments*, New York: Hafner Publishing Company, reprint of the 8th ed.

Fitzpatrick, M. D. (2010): “Preschoolers Enrolled and Mothers at Work? The Effects of Universal Prekindergarten,” *Journal of Labor Economics*, 28, 51–85.

——— (2012): “Revising Our Thinking About the Relationship Between Maternal Labor Supply and Preschool,” *Journal of Human Resources*, 47, 583–612.

Foster, G. (2006): “It’s not your peers, and it’s not your friends: Some progress toward understanding the educational peer effect mechanism,” *Journal of Public Economics*, 90, 1455–1475.

Frank, R. G. and K. Lamiraud (2009): “Choice, price competition and complexity in markets for health insurance,” *Journal of Economic Behavior & Organization*, 71, 550–562.

Frölich, M. (2006): “Non-parametric regression for binary dependent variables,” *Econometrics Journal*, 9, 511–540.

——— (2007): “Nonparametric IV estimation of local average treatment effects with covariates,” *Journal of Econometrics*, 139, 35–75.

Frölich, M. and M. Lechner (2010): “Exploiting Regional Treatment Intensity for the Evaluation of Labor Market Policies,” *Journal of the American Statistical Association*, 105, 1014–1029.

FSIO (2010): “Financial Support for Extra-Familiar Childcare: Results after 7 Years,” `http://www.bsv.admin.ch/praxis/kinderbetreuung/00112/`, retrieved March 31, 2014.

FSO (2004): “Einkommens- und Verbrauchserhebung 2004,” `http://www.admin.ch/00090/index.html?lang=de&msg-id=7358`, retrieved April 3, 2014.

Gelbach, J. B. (2002): “Public Schooling for Young Children and Maternal Labor Supply,” *American Economic Review*, 92, 307–322.

Gettens, J. W. (2009): "Medicaid expansions: The work and program participation of people with disabilities," Ph.D. thesis, Brandais University.

Goux, D. and E. Maurin (2010): "Public school availability for two-year olds and mothers' labour supply," *Labour Economics*, 17, 951–962.

Graham, B. S. (2011): "Chapter 19 - Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers," North-Holland, vol. 1 of *Handbook of Social Economics*, 965–1052.

Graham, B. S., G. W. Imbens, and G. Ridder (2010): "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach," Working Paper 16499, National Bureau of Economic Research.

Greene, J. and M. Winters (2007): "Revisiting grade retention: An evaluation of Florida's test-based promotion policy," *Education Finance and Policy*, 2, 319–340.

Havnes, T. and M. Mogstad (2011): "Money for nothing? Universal child care and maternal employment," *Journal of Public Economics*, 95, 1455–1465.

Holmes, T. J. (1998): "The Effect of State Policies on the Location of Manufacturing: Evidence from State Borders," *Journal of Political Economy*, 106, 667–705.

Hoxby, C. (2000): "Peer Effects in the Classroom: Learning from Gender and Race Variation," Working Paper 7867, National Bureau of Economic Research.

Huber, M., M. Lechner, and C. Wunsch (2013): "The performance of estimators based on the propensity score," *Journal of Econometrics*, 175, 1–21.

Imbens, G. and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–75.

JACOB, B. A. AND L. LEFGREN (2004): "Remedial education and student achievement: A regression-discontinuity analysis," *Review of economics and statistics*, 86, 226–244.

——— (2009): "The Effect of Grade Retention on High School Completion," *American Economic Journal: Applied Economics*, 1, 33–58.

JIMERSON, S. (2001): "Meta-analysis of grade retention research: Implications for practice in the 21st century," *School Psychology Review*, 30, 420–437.

KAMETTE, F. (2011): "Organisation of School Time in the European Union," Policy Paper 212, Fondation Robert Schumann.

KIMMEL, J. (1998): "Child Care Costs As A Barrier To Employment For Single And Married Mothers," *The Review of Economics and Statistics*, 80, 287–299.

KORNFELD, R. AND K. RUPP (2000): "Net Effects of the Project NetWork Return-to-Work Case Management Experiement on Participant Earnings, Benefit Receipt, and Other Outcomes," *Social Security Bulletin*, 63, 12–33.

KOSTØL, A. R. AND M. MOGSTAD (2014): "How Financial Incentives Induce Disability Insurance Recipients to Return to Work," *American Economic Review*, 104, 624–55.

KREMER, M., E. DUFLO, AND P. DUPAS (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking," *American Economic Review*, 101, 1739 –1774.

KREMER, M. AND D. LEVY (2008): "Peer Effects and Alcohol Use among College Students," *The Journal of Economic Perspectives*, 22, 189–206.

LAVY, V., M. D. PASERMAN, AND A. SCHLOSSER (2008): "Inside the Black of Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom," Working Paper 14415, National Bureau of Economic Research.

LAZEAR, E. P. (2001): "Educational Production," *The Quarterly Journal of Economics*, 116, 777–803.

Lee, D. S. and T. Lemieux (2010): “Regression Discontinuity Designs in Economics,” *Journal of Economic Literature*, 48, 281–355.

Lefebvre, P. and P. Merrigan (2008): “Child-Care Policy and the Labor Supply of Mothers with Young Children: A Natural Experiment from Canada,” *Journal of Labor Economics*, 26, 519–548.

Lorence, J. and A. Dworkin (2006): “Elementary grade retention in Texas and reading achievement among racial groups: 1994–2002,” *Review of Policy Research*, 23, 999–1033.

Low, H. and L. Pistaferri (2010): “Disability Risk, Disability Insurance and Life Cycle Behavior,” Working Paper 15962, National Bureau of Economic Research.

Lundin, D., E. Mörk, and B. Öckert (2008): “How far can reduced childcare prices push female labour supply?” *Labour Economics*, 15, 647–659.

Lyle, D. S. (2007): “Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point,” *The Review of Economics and Statistics*, 89, 289–299.

——— (2009): “The Effects of Peer Group Heterogeneity on the Production of Human Capital at West Point,” *American Economic Journal: Applied Economics*, 1, 69–84.

Manacorda, M. (2012): “The Cost of Grade Retention,” *The Review of Economics and Statistics*, 94, 596–606.

Manski, C. F. (1989): “Schooling as experimentation: a reappraisal of the post-secondary dropout phenomenon,” *Economics of Education Review*, 8, 305–312.

——— (1993): “Identification of Endogenous Social Effects: The Reflection Problem,” *Review of Economic Studies*, 60, 531–42.

——— (2004): “Statistical Treatment Rules for Heterogeneous Populations,” *Econometrica*, 72, 1221–1246.

MARIE, O. AND J. VALL CASTELLO (2012): "Measuring the (income) effect of disability insurance generosity on labour market participation," *Journal of Public Economics*, 96, 198–210.

MCCRARY, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142, 698–714.

MICHALOPOULOS, C., P. K. ROBINS, AND I. GARFINKEL (1992): "A Structural Model of Labor Supply and Child Care Demand," *Journal of Human Resources*, 27, 166–203.

MORETTI, E. (2004): "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions," *American Economic Review*, 94, 656–690.

NIEDERLE, M. AND L. VESTERLUND (2007): "Do Women Shy Away From Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics*, 122, 1067–1101.

NOLLENBERGER, N. AND N. RODRÍGUEZ-PLANAS (2011): "Child Care, Maternal Employment and Persistence: A Natural Experiment from Spain," IZA Discussion Papers 5888, Institute for the Study of Labor (IZA).

OECD (2001): *OECD Employment Outlook*, Paris: OECD Publishing.

——— (2003): *Transforming Disability into Ability*, Paris: OECD Publishing.

——— (2008): *Higher Education to 2030*, vol. 1, Demography, Paris: OECD Publishing.

——— (2009): *Sickness, Disability and Work: Keeping on track in the economic downturn*, Stockholm: OECD Publishing.

——— (2010): *Sickness, Disability and Work: Breaking the Barriers - A Synthesis of Findings across OECD Countries*, Paris: OECD Publishing.

——— (2012): "OECD Family Database," `http://www.oecd.org/social/family/database`, retrieved September 9, 2013.

Oosterbeek, H. and R. van Ewijk (2014): "Gender peer effects in university: Evidence from a randomized experiment," *Economics of Education Review*, 38, 51–63.

Ou, D. (2010): "To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam," *Economics of Education Review*, 29, 171–186, special Issue in Honor of Henry M. Levin.

Pence, K. M. (2006): "Foreclosing on Opportunity: State Laws and Mortgage Credit," *The Review of Economics and Statistics*, 88, 177–182.

Pinto, C. (2011): "Semiparametric Estimation of Peer Effects in Classrooms," Mimeo.

Roderick, M. and J. Nagaoka (2005): "Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless?" *Educational Evaluation and Policy Analysis*, 27, 309.

Sacerdote, B. (2001): "Peer Effects With Random Assignment: Results For Dartmouth Roommates," *The Quarterly Journal of Economics*, 116, 681–704.

——— (2011): *Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?*, Elsevier, vol. 3 of *Handbook of the Economics of Education*, chap. 4, 249–277.

Samuelson, W. and R. Zeckhauser (1988): "Status Quo Bias in Decision Making," *Journal of Risk and Uncertainty*, 1, 7–59.

Schlosser, A. (2011): "Public Preschool and the Labor Supply of Arab Mothers: Evidence from a Natural Experiment," Mimeo.

Stapleton, D., G. Livermore, C. Thornton, B. O'Day, R. Weathers, K. Harrison, S. O'Neil, E. S. Martin, D. Wittenburg, and D. Wright

(2008): "Ticket to Work at the Crossroads: A Solid Foundation with an Uncertain Future," Mathematica Policy Research Reports 6153, Mathematica Policy Research.

Staubli, S. (2011): "The impact of stricter criteria for disability insurance on labor force participation," *Journal of Public Economics*, 95, 1223–1235.

Stinebrickner, R. and T. R. Stinebrickner (2006): "What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds," *Journal of Public Economics*, 90, 1435–1454.

Thornton, C., G. Livermore, D. Stapleton, J. Kregel, T. Silva, B. O'Day, T. Fraker, W. G. R. Jr., H. Schroeder, and M. Edwards (2004): "Evaluation of the Ticket to Work Program Initial Evaluation Report," Mathematica Policy Research Reports 4154, Mathematica Policy Research.

Tinto, V. (1975): "Dropout from Higher Education: A Theoretical Synthesis of Recent Research," *Review of Educational Research*, 45, 89–125.

Vuren, A. and D. Vuuren (2007): "Financial Incentives in Disability Insurance in the Netherlands," *De Economist*, 155, 73–98.

Waldfogel, J. (1997): "The effect of children on women's wages," *American Sociological Review*, 62, 209–217.

Weathers, R. R. and J. Hemmeter (2011): "The impact of changing financial work incentives on the earnings of Social Security Disability Insurance (SSDI) beneficiaries," *Journal of Policy Analysis and Management*, 30, 708–728.

Whitmore, D. (2005): "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment," *American Economic Review*, 95, 199–203.

Wooldridge, J. M. (2005): "Unobserved heterogeneity and estimation of average partial effects," in *Identification and Inference for Econometric Models: Essays in*

Honor of Thomas Rothenberg, ed. by J. S. D.W.K. Andrews, Cambridge University Press, 27–55.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge: The MIT Press, 2nd ed.

Zimmerman, D. J. (2003): "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment," *The Review of Economics and Statistics*, 85, 9–23.

Curriculum Vitae

Education

2009 – 2015	PhD, Economics, University of St. Gallen Program: PhD in Economics and Finance (PEF) Main adviser: Prof. Dr. Michael Lechner
2013 – 2014	Visiting Student Researcher, Economics Department, UC Berkeley Funded by the Swiss National Science Foundation Sponsor: Prof. Bryan S. Graham, PhD
2009 – 2010	Swiss Program for Beginning Doctoral Students in Economics Study Center Gerzensee, Foundation of the Swiss National Bank
2007 – 2009	M.A., Economics, University of St. Gallen
2000 – 2006	Diploma (1. Staatsexamen), Music and German, University of Cologne

Professional experience

Since 2014	Postdoctoral Research Associate, University of Southern California USC Dornsife Institute for New Economic Thinking
2009 – 2012	Research assistant, University of St. Gallen Swiss Insitute for Empirical Economic Research Chair Prof. Dr. Michael Lechner
2007 – 2009	Student research assistant, University of St. Gallen Swiss Insitute for Empirical Economic Research Chair Prof. Dr. Michael Lechner