Educational Reforms, Incentive Schemes, and Their Impact on Academic Achievement

Dissertation

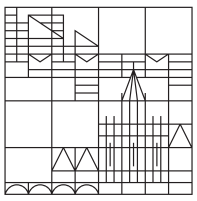
zur Erlangung des akademischen Grades eines Doktors der

Wirtschaftswissenschaften (Dr.rer.pol.)

vorgelegt von

Michael Dörsam

Universität Konstanz



Sektion Politik - Recht - Wirtschaft

Fachbereich Wirtschaftswisssenschaften

Konstanz, 2018

Tag der mündlichen Prüfung: 7. Mai 2018

Erster Referent: Prof. Dr. Guido Schwerdt

Zweiter Referent: Prof. Dr. Friedrich Breyer

Prüfungsvorsitzender: Jun.-Prof. Dr. Stephan Maurer

Danksagung

Ich möchte mich an dieser Stelle bei allen Menschen bedanken, die zum Gelingen dieser Arbeit beigetragen haben:

Besonders bedanke ich mich bei meinem Erstbetreuer, Herrn Professor Dr. Guido Schwerdt, der mir zu jeder Zeit mit wertvollem Rat zur Seite stand. Sein Feedback hat mir stets bei der Formulierung von Forschungsfragen und deren Durchführung geholfen und damit maßgeblich zum Erfolg dieser Dissertation beigetragen.

Ebenso möchte ich mich bei Herrn Professor Dr. Friedrich Breyer für seine Bereitschaft bedanken, die Zweitbetreuung und Begutachtung meiner Dissertation übernommen zu haben sowie bei Frau Prof. Dr. Anja Schöttner, die meine Promotion insbesondere während meiner Anfangszeit in Konstanz begleitet hat.

Ferner bedanke ich mich bei Herrn Manfred Witznick und Herrn Matthias Moebius, den Mitarbeitern des Forschungsdatenzentrums sowie den vielen wissenschaftlichen Hilfskräften, die durch Ihren Einsatz meine Dissertation in dieser Form erst ermöglichten. Für die Unterstützung bei allen administrativen Angelegenheiten danke ich Frau Gundula Hadjiani und Frau Heike Knappe. Meinen Freunden und Kollegen möchte ich für die schöne Zeit danken, die ich dank Ihnen innerhalb und außerhalb der Universität hatte und durch die ich immer wieder mit neuer Energie ans Werk gehen konnte.

Großen Einfluss auf meinen eingeschlagenen Weg hatten auch Frau Dr. Iris Claus, Frau Dr. Susanne Neckermann, Herr Prof. Dr. Thorsten Schank sowie Herr Andrew Whiteford, die mir Ihr Vertrauen schenkten und auf diese Weise mein Interesse am wissenschaftlichen Arbeiten weckten.

Mein ganz besonderer Dank gilt meiner Familie und meiner Partnerin, ohne deren Unterstützung und gewährter Freiheit mein bisheriger Werdegang nicht möglich gewesen wäre und denen ich diese Arbeit widme.

Contents

List of Figures

List of Tables

Zusammenfassung

Die vorliegende Dissertation besteht aus drei unabhängigen Forschungspapieren, die in den Jahren 2013 bis 2017 während meiner Zeit an der Universität Konstanz als Teilnehmer am "Doctoral Programme in Quantitative Economics and Finance" entstanden sind. Obwohl die Kapitel in sich geschlossen sind, haben sie einen gemeinsamen Fokus auf die Evaluierung von Reformen und Charakteristika des deutschen Bildungssystem. Durch die Nutzung verschiedener Datenquellen und die Anwendung fortgeschrittener ökonometrischer Methoden, untersucht die Doktorarbeit empirisch die Effekte dieser Reformen und Charakteristika des Bildungssystems auf Schüler- und Studierendenleistungen. Das erste Kapitel befasst sich mit der Anreizwirkung des gegliederten Schulsystems auf das Anstrengungsniveau von Grundschülern vor dem Übertritt in die weiterführende Schule. Das zweite Kapitel untersucht den Effekt der Verkürzung der Gymnasialzeit um ein Jahr (bei gleichzeitiger Beibehaltung des Lehrplans und der Jahreswochenstunden) auf die Leistung von Studenten. Das dritte Kapitel widmet sich dem Effekt der Einführung sogenannter Orientierungsprüfungen an den Universitäten in Baden-Württemberg im Jahr 2001 auf die Abbruch- und Abschlusswahrscheinlichkeit der Studierenden.

Im ersten Kapitel meiner Dissertation liefere ich neue Erkenntnisse zur Anreizwirkung eines gegliederten Schulsystems auf das Anstrengungsniveau und die Kompetenzentwicklung von Grundschülern vor dem Übertritt in die weiterführende Schule. Viele theoretische und empirische Studien aus dem Bereich der Arbeitsmarkt- und Experimentalökonomie zeigen, das Turniere in einem Beschäftigungskontext zu höherer Anstrengung führen können (für einen Überblick, siehe Prendergast, 1999; Dechenaux et al., 2015). Nur wenige Studien haben sich jedoch bislang mit der Turnierstruktur beschäftigt, welche durch ein gegliedertes Schulsystem besteht. Da der Besuch einer höheren Schulform der Sekundarstufe I mit signifikant besseren Karriereperspektiven verbunden ist, sollten Schüler, die sich an der Grenze zur Zulassung zur nächsthöheren Schulform befinden, ihre Anstrengung im Jahr vor der Schulformempfehlung relativ mehr erhöhen als Schüler, die sich nicht an der Grenze befinden. In dem ich Variation in der Anreizintensität zwischen solch marginalen und nicht-marginalen Schülern in zwei Bundesländern über die Zeit nutze, finde ich Hinweise, welche diese theoretische Vorhersage bestätigen. Das gestiegene

Anstrengungsniveau marginaler Schüler geht dabei mit einer relativ größeren Verbesserung in standardisierten Kompetenzmaßen einher, ohne gleichzeitig einen negativen Effekt auf die Lernfreude zu haben.

Der Beitrag dieses Kapitels zur Literatur ist dreifach: Erstens liefern meine Ergebnisse einen empirischen Beleg für das theoretische Modell von Eisenkopf (2009), welches prognostiziert, dass sich Grundschüler mehr anstrengen, falls sie, basierend auf ihren schulischen Leistungen, beim Übergang in die Sekundarstufe unterschiedlichen Schulformen zugeordnet werden. Die Studie von Koerselman (2013) ist bislang die einzige, die empirisch den Effekt eines gegliederten Schulsystems auf Schülerleistungen vor dem Übergang untersucht. Koerselman (2013) beobachtet das Anstrengungsniveau der Grundschüler jedoch nicht direkt, sondern nur indirekt über schulische Testergebnisse, und berichtet darüber hinaus von zwei Selektionsproblemen, welche die Schätzer möglicherweise verzerren. Zweitens zeigen meine Ergebnisse, dass Anreize, die durch Richtlinien in früheren Jahren der Schulbildung entstehen, bei der Untersuchung von schulischen Maßnahmen in späteren Jahren berücksichtigt werden müssen. Drittens hat meine Analyse methodische Implikationen. "Mehrwertmodelle" beispielsweise, welche frühere Testergebnisse zur Kontrolle nicht beobachtbarer Fähigkeiten nutzen, könnten verzerrte Schätzer liefern, falls die früheren Ergebnisse bereits durch die betrachtete Maßnahme beeinflusst waren. Die Existenz von Anreizeffekten könnte ebenso die Nutzung früherer Ergebnisse für Placebo-Tests invalidieren. In einem sorgfältig kontrollierten Experiment gibt es vor dem Treatment keinen Unterschied in den beobachtbaren Leistungen zwischen der Behandlungs- und der Kontrollgruppe. Im Falle von natürlichen Experimenten kann es jedoch sein, dass den Beobachtungssubjekten ihr zukünftiger Behandlungsstatus bewusst ist und sie sich entsprechend verhalten. Pischke and Manning (2006) finden beispielsweise einen Zusammenhang zwischen dem Anstieg von Testergebnissen im Alter von 7 bis 11 und der Existenz eines gegliederten Schulsystems in Großbritannien.

Das zweite Kapitel entstand durch die gemeinsame Arbeit mit Frau Dr. Verena Lauber, einer ehemaligen Kommilitonin im "Doctoral Programme in Quantitative Economics and Finance". Wir untersuchen in diesem Papier eine aktuelle Reform des deutschen Bildungssystems, wodurch die Gymnasialzeit um ein Jahr gekürzt wurde, der Lehrplan und die Jahreswochenstundenzahl jedoch beibehalten wurden. Die Reform ermöglicht es uns somit, neue Erkenntnisse zu dem Zusammenhang zwischen dem Input 'Anzahl Schuljahre' – einem der am häufigsten diskutierten Faktoren in der Bildungsproduktionsfunktion – und dem Output 'Schülerleistung' zu gewinnen. Auf der einen Seite zeigen mehrere Studien, dass die Anzahl der Schuljahre sowie die Unterrichtszeit positiv mit Bildungserfolg sowie monetären und nicht monetären Leistungen korrelieren (siehe z.B. Bellei, 2009; Wößmann, 2003; Card, 1999; Lochner, 2011). Auf der anderen Seite wird der Arbeitsmarkteintritt

durch eine Verlängerung der Schulzeit verzögert und die Dauer der Erwerbstätigkeit verkürzt. Hanushek and Wößmann (2008) zeigen darüber hinaus, dass es eher kognitive Fähigkeiten denn reiner Schulerfolg sind, die ökonomisches Wohlergehen bestimmen, und dass die Qualität der schulischen Einrichtungen entscheidend ist. Wir nutzen einen einzigartigen Paneldatensatz, um den Effekt der Schulzeitverkürzung auf die Leistung von Studenten zu untersuchen. Durch Variation in der Implementierung der Reform zwischen Schulformen über die Zeit können wir den Reformeffekt von Kohorten-, Bundesland- und Schulformeffekten trennen. Unsere Schätzungen legen nahe, dass die Reform die Opportunitätskosten der schulischen Bildung senkt und einen früheren Arbeitsmarkteintritt erleichtert, da wir keine schädlichen Effekte finden, während die Studenten durchschnittlich ein Jahr jünger sind.

Das dritte Kapitel ist in Zusammenarbeit mit Herrn Enzo Brox, einem Kommilitonen aus der Graduiertenschule für Entscheidungswissenschaften an der Universität Konstanz entstanden. In diesem Papier liefern wir erste Hinweise zu den Effekten einer High-Stakes-Testing Regelung, welche an Universitäten in Baden-Württemberg implementiert wurde. Im September 1999 verabschiedete die baden-württembergische Landesregierung ein Gesetz, welches Universitäten dazu verpflichtete, zum Wintersemester 2000/01 in jedem Studiengang so genannte Orientierungsprüfungen einzuführen. Zu diesem Zeitpunkt hatte kein anderes Bundesland eine vergleichbare High-Stakes-Testing Regelung implementiert, wohingegen ähnliche Regelungen unter anderem an niederländischen und Schweizer Universitäten zu finden sind (see Vossensteyn et al., 2015). Das Ziel einer solchen Regelung ist es, bereits zu einem frühen Zeitpunkt zu überprüfen, ob ein Student die nötigen Fähigkeiten für einen Studienabschluss mitbringt, und jene Studenten, die die Anforderungen nicht erfüllen, frühzeitig auszuschließen. Die Effektivität einer solchen Maßnahme wurde jedoch bislang kaum untersucht.

Im ersten Teil des Papiers nutzen wir administrative Paneldaten, aggregiert auf Studiengangsebene, über alle Studenten, die in Deutschland zwischen 1997 und 2003 an einer Universität eingeschrieben waren, um den durchschnittlichen kausalen Effekt der High-Stakes-Testing Regelung auf Studienabbruch innerhalb der ersten beiden Jahre zu schätzen. Da Orientierungsprüfungen zu diesem Zeitpunkt nur an baden-württembergischen Universitäten implementiert waren, können wir den Effekt von Kohorten-, Bundesland-, Universitäts- und Studiengangseffekten durch die Schätzung von Differenz-in-Differenzen (DiD) Modellen trennen. Dadurch gewinnen wir eine erste Vorstellung von der Wirksamkeit der Regelung. Im zweiten Teil des Papiers wenden wir eine "scharfe" Regressions-Diskontinuitäts-Analyse (RD Analyse) auf administrative, studentische Paneledaten des Bachelor-Studiengangs Wirtschaftswissenschaften an der Universität Konstanz an. Dies ermöglicht es uns, den Mechanismus der High-Stakes-Testing Regelung genauer

zu untersuchen, da wir Studenten und ihre Studienleistungen bis zum Abschluss bzw. ihrer Exmatrikulation beobachten.

Unsere DiD-Schätzungen zeigen, dass die Einführung von Orientierungsprüfungen in Baden-Württemberg die durchschnittliche Abbruchrate nach zwei Jahren um drei Prozentpunkte erhöht hat, was einem Anstieg von 10 Prozent entspricht. Die RD Analyse im zweiten Teil des Papiers zeigt, dass Studienanfänger, die Pech hatten und eine Orientierungsprüfung knapp nicht bestanden haben, eine um 16 bis 19 Prozentpunkte höhere Wahrscheinlichkeit aufweisen, nach dem ersten Semester auszuscheiden, als Studenten, die dieselbe Orientierungsprüfung knapp bestanden haben. Dieser Effekt verringert sich über die Zeit, bleibt aber bis zum Ende des Studiums bestehen, da wir immer noch einen Unterschied in der Abschlusswahrscheinlichkeit von 10 bis 13 Prozentpunkte finden. Unsere Studie verdeutlicht somit die Schwierigkeit der Implementierung dieser High-Stakes-Testing Regelung, welche im Finden des optimalen Bewertungsstandards liegt. Trotz nahezu identischer Fähigkeiten können Studenten, die die erforderliche Punktzahl in einer Orientierungsprüfung knapp nicht erreichen, eine geringere Wahrscheinlichkeit haben ihren Abschluss zu machen, als Studenten, die die erforderliche Punktzahl knapp erreichen.

Summary

This dissertation consists of three stand-alone research papers that were written in the years 2013 to 2017 during my doctoral studies at the University of Konstanz as a participant of the "Doctoral Programme in Quantitative Economics and Finance". Though these chapters are self-contained, they have a common focus on the evaluation of education policies implemented in the German education system at the primary, secondary, and tertiary level. Based on unique data sources and advanced microeconometric methods, this thesis empirically investigates the effects of these reforms on different student outcomes. The first chapter deals with the incentive effect of educational tracking on the effort provision of primary school students before the track decision is made. The second chapter investigates the effect of a high school reform that reduced the years of schooling by one year but left the curriculum, and the total class time unchanged, on university students' academic achievement. The third chapter investigates the impact of a high-stakes testing policy implemented at state universities in Baden-Württemberg on drop-out on graduation rates.

In the first chapter of this thesis, I provide novel insights on the incentive effects of educational tracking on student outcomes measured prior to the track decision. While many theoretical and experimental studies in the field of labor and experimental economics show that tournaments lead to increased efforts in employment relationships (for an overview, see Prendergast, 1999; Dechenaux et al., 2015), only very few studies address the tournament structure provided by educational tracking. As the attendance of a higher ability track is associated with significantly higher career prospects, students who are at the margin to be admitted to a higher track should increase their effort prior to the admittance decision relatively stronger than students who are above the margin. Exploiting variation in the incentive intensity between marginal and non-marginal students in two German states over time, I find evidence confirming this theoretical prediction. The increased effort of marginal students is accompanied by a relatively greater improvement in standardized competence measures without having a detrimental effect on joy of learning.

The contribution of this chapter to the literature is threefold: First, my findings provide empirical evidence for the theoretical model by Eisenkopf (2009) which predicts

that students exert more effort in elementary school if they are separated into different secondary schools based on their observed academic performance. So far, only Koerselman (2013) investigated empirically the effects of tracking on student outcomes measured prior to the track decision. However, he does not observe student effort directly but only indirectly via achievement test scores, and further reports two selection problems that potentially bias his estimates. Second, my results show that incentives from policies in earlier years of education must be taken into account when investigating the effect of policies at later ages. Third, there are methodological implications. For example, value added models using test scores to control for unobservables may provide biased estimates if the early age outcomes are affected by the policy under consideration. The existence of incentive effects may also invalidate the use of early outcomes in placebo tests. In a carefully controlled experiment, there is no difference in pre-treatment outcomes between the treatment and the control group. In natural experiments, however, subjects may be aware of their future treatment status and behave accordingly. For example, Pischke and Manning (2006) find that test score growth in the UK between age 7 and 11 is correlated with tracking policies after the age of 11.

The second chapter of this thesis is joint work with Dr. Verena Lauber, a former fellow student in the "Doctoral Programme in Quantitative Economics and Finance". In this paper, we investigate a recent education reform in Germany that reduced the duration of academic high school education by one year but left the curriculum, and total class time unchanged. The reform thus allows us to provide new insights on the relation between the input 'years of schooling' – one of the most frequently discussed factors in the education production function – and the output 'student achievement'. On the one hand, several studies show that instruction time and years of schooling are positively related to academic achievement as well as monetary and non-monetary benefits (see, e.g., Bellei, 2009; Wößmann, 2003; Card, 1999; Lochner, 2011). On the other hand, the entry into the labor force is delayed and the duration of gainful employment reduced with an increasing length of schooling. Hanushek and Wößmann (2008) further show that cognitive skills rather than mere school attainment determine economic well-being, and that the quality of school institutions is decisive. We use a unique data set of university students to investigate the effect of the one-year reduction in years of schooling on academic achievement at the tertiary level. By exploiting variation in the implementation of the reform across school types over time, we isolate the reform effect from cohort, state, and school type effects. Our estimates suggest that the reform lowers the opportunity costs of schooling and facilitates an earlier labor market entry as we find no detrimental effects while students are one year younger on average.

The third chapter of this thesis is joint work with Enzo Brox, a fellow student at

the Graduate School of Decision Sciences at the University of Konstanz. In this paper, we provide novel evidence on the effect of a high-stakes testing policy implemented at universities in a large federal state in Germany. In September 1999, the government of Baden-Württemberg passed a bill that obliged universities to introduce so-called orientation exams in each study program from the winter term 2000/01 onward. At this time, no other federal state had implemented a comparable high-stakes testing policy, whereas similar policies are implemented among others at Dutch and Swiss universities (see Vossensteyn et al., 2015). The aim of such a policy is to test at an early stage whether a student has the required skills to graduate, and to exclude those students who do not satisfy the requirements as early as possible. The effectiveness of such a policy, however, has been barely investigated so far.

In the first part of this paper, we use administrative panel data, aggregated at the study program level[1], on all university students in Germany from 1997 to 2003 to estimate the average causal effect of this high-stakes testing policy on drop-out within the first two years. As orientation exams were only introduced at universities in Baden-Württemberg, we can disentangle the effect of orientation exams from cohort, state, university, and study program fixed effects by estimating a difference-in-differences (DiD) model. This allows us to gain a first idea of the effectiveness of the policy. In the second part of this paper, we apply a sharp regression discontinuity design (RDD) to administrative, student-level panel data of the Economics program at the University of Konstanz. This enables us to investigate in more detail the mechanism of the high-stakes testing policy, as we observe students and their academic achievement up to graduation (or drop-out, respectively).

Our DiD estimates show that the introduction of orientation exams in Baden-Württemberg increased average drop-out rates after two years by about three percentage points, which is equivalent to an increase of about ten percent. The analysis in the second part of this paper reveals that freshmen students who have bad luck and narrowly fail an orientation exam are by about 16 to 19 percentage points more likely to drop out after the first semester. This effect diminishes over time but persists until the end of their studies, as we still find a 10 to 13 percentage points difference in the probability to graduate. Thus, our study highlights the difficulty in implementing this high-stakes testing policy, which is to find the optimal grading standard. Although having approximately the same ability, students who just fail to achieve the required threshold in an orientation exam have a lower probability to graduate than students who just pass the threshold.

[1] A study program refers to one specific program, at one specific university.

References

Bellei, C. (2009). Does lengthening the school day increase students academic achievement? Results from a natural experiment in Chile. *Economics of Education Review*, 28(5):629–640.

Card, D. (1999). The Causal Effect of Education on Earnings. In Ashenfelter, O. C. and Card, D., editors, *Handbook of Labor Economics*, volume 3, pages 1801 – 1863. Amsterdam: Elsevier.

Dechenaux, E., Kovenock, D., and Sheremeta, R. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

Eisenkopf, G. (2009). Student selection and incentives. *Zeitschrift für Betriebswirtschaft*, 79(5):563–577.

Hanushek, E. A. and Wößmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3):607–68.

Koerselman, K. (2013). Incentives from curriculum tracking. *Economics of Education Review*, 32:140–150.

Lochner, L. (2011). Nonproduction benefits of education: Crime, health, and good citizenship. *Handbook of the Economics of Education*, 4:183.

Pischke, J.-S. and Manning, A. (2006). Comprehensive versus selective schooling in England in Wales: What do we know? Working Paper 12176, National Bureau of Economic Research.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63.

Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., and Wollscheid, S. (2015). Dropout and completion in higher education in Europe: Main report. Technical report, Luxembourg: Publications Office of the European Union.

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2):117–170.

Chapter 1

The Incentive Effect of Tracking on Student Effort and Skill Development

1.1 Introduction

The seminal paper by Lazear and Rosen (1981) shows that tournaments lead to increased efforts in employment relationships. By offering a prize to the winner, tournaments induce agents to increase their efforts as long as the agents' benefits from additional output exceed their effort costs.[1] This theoretical prediction has been confirmed by many empirical and experimental studies in the field of labor economics (for an overview, see Prendergast, 1999; Dechenaux et al., 2015). A similar incentive scheme is present in education systems that separate students by academic ability at a certain stage. As the attendance of a higher ability track is associated with significantly higher career prospects, students have a strong incentive to get into the highest possible track. In a theoretical model, Eisenkopf (2009) shows that students exert more effort in elementary school if they are separated into different secondary schools based on their observed academic performance. Empirical evidence for this incentive scheme, however, is scarce.

This paper helps to fill this gap by investigating the incentive effects of tracking prior to the track decision. In particular, I test the following four hypotheses: First, students who are – based on their academic performance – at the margin to be admitted to a higher track (marginal students), increase their effort prior to the admittance decision relatively more than students who are above the margin (non-marginal students). Conversely, the effort of marginal students decreases relative to the effort of non-marginal students once the track decision is made. Second, strict admission policies for attending a certain track are crucial for the existence of an incentive effect of tracking. Third, marginal students experience a relative decline in joy of learning compared to non-marginal students. And fourth, the increased efforts of marginal students lead to a relatively stronger improvement in skills, measured by standardized competence tests.

The analysis is based on data of the BiKS-8-14 study ('Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vor- und Grundschulalter'), a large individual-level panel study of elementary school students, their class teachers, and their parents. The rich data set allows me to compare the outcomes of marginal and non-marginal students over time, and additionally provides me with variation in institutional admission policies for attending a certain track. While in Bavaria, the secondary school track to which parents can send their child depends on the child's grade point average (GPA) in fourth grade, the secondary school track choice in Hesse is independent of a child's GPA. I investigate this variation in the incentive intensity between marginal and

[1] In more complex and creative task environments, the effects become more ambiguous, for example, because of detrimental effects on intrinsic motivation (see, e.g., Amabile, 1996; Frey and Oberholzer-Gee, 1997; Frey and Jegen, 2001), or because of inequity or loss aversion (see, e.g., Eisenkopf and Teyssier, 2013).

non-marginal students from Bavaria and Hesse in difference-in-differences (DiD) models.[2] The identifying assumption is that there are no other factors explaining a different effort provision of marginal and non-marginal students in third and fourth grade except for the status of being at the margin to a higher track.

In line with the theoretical predictions, I find for Bavarian students a significant positive effect of being at the margin to the high school track on effort in the school term before the track decision is made. The different career prospects of the highest and the intermediate track in conjunction with the strict admission policy induce the marginal Bavarian students to increase their effort in fourth grade relatively more than the non-marginal Bavarian students. In contrast, I find no significant positive effect for the marginal students from Hesse. Once the decision is made, marginal Bavarian students slightly decrease their efforts relative to non-marginal students. At the same time, the increase in effort of the marginal Bavarian students is accompanied by a relatively stronger improvement in standardized competence measures without having a detrimental effect on joy of learning.

The key contribution of this paper is to test whether tracking works as an incentive scheme prior to the track decision. While many studies investigated the effect of tracking on student outcomes measured after the track decision (see, e.g., Meghir and Palme, 2005; Hanushek and Wößmann, 2006; Duflo et al., 2011), only Koerselman (2013) investigated the effects of tracking on student outcomes measured prior to the track decision. Considering a school reform conducted in the UK in the 1960s and 70s which gradually implemented a comprehensive school system, his estimates from multilevel linear models suggest that tracking has a positive effect on student effort in the years before the track decision is made. However, he does not observe student effort directly but only indirectly via achievement test scores. Furthermore, he reports two selection problems: First, the reform was not implemented randomly across regions but richer, right-wing areas implemented the reform more slowly. Second, parents of good students had an incentive to move to a tracked area while parents of poor students did not.

The BiKS-8-14 data allow to investigate the effect of tracking on student effort directly. In addition to the effort measure assessed by the students and the class teachers, the data include standardized competence measures in reading, writing, and analytic thinking, a measure of joy of learning, information on grades, teacher recommendations for secondary school tracks, as well as a large set of family background variables. Exploiting the panel structure of the data, I control for unobserved but fixed heterogeneity across students and teachers. The identification strategy does furthermore not suffer from the threats discussed by Koerselman (2013) as, first, the admission policies in Bavaria and Hesse had

[2]The official threshold for high school entrance in Bavaria is at 2.33. The empirical threshold in the data for high school entrance in Hesse is also at 2.33. I define a student as a marginal student if his GPA at the end of third grade is at 2.0 to 2.67.

been implemented long before the students of the sample enrolled in elementary school, and second, there is no evidence for selective migration due to the policies.

My results show that higher effort and student achievement may not only be achieved in secondary school through minimum competency tests or external exit exams, but also in elementary school through tracking. The results further show that incentives from policies in earlier years of education must be taken into account when investigating the effect of policies at later ages. For value added specifications that use early test scores to control for unobservables my results imply that estimates are biased if the early age outcomes are affected by the policy under consideration. The existence of incentive effects may also invalidate the use of early outcomes in placebo tests. While in a carefully controlled experiment, there is no difference in pre-treatment outcomes between the treatment and the control group, in natural experiments subjects may be aware of their future treatment status and behave accordingly.

The paper proceeds as follows: In section 1.2, I discuss the theoretical framework and derive the research hypotheses. Section 1.3 presents background information on the German education system, and the tracking policies of Bavaria and Hesse. Section 1.4 describes the data. In section 1.5, I present the empirical strategy. Section 1.6 presents the results. Section 1.7 discusses the results and concludes.

1.2 Theoretical Framework

In this section, I discuss the related theoretical literature and derive the research hypotheses. Tournaments in an economic context can be described as follows: A group of agents competes for a fixed set of prizes which are specified by a principal in advance. To increase their likelihood of winning a better prize, agents exert effort. As in a sports game, the winner of a tournament is determined by his relative level of performance, i.e., his rank.

In the context of optimum labor contracts, Lazear and Rosen (1981) have shown that rank-order tournaments efficiently incentivize risk-neutral workers to exert effort, and to sort workers according to their skills.[3] This theoretical model has been confirmed by many empirical and experimental studies in the field of labor economics (for an overview, see Prendergast, 1999; Dechenaux et al., 2015).

Related to this work is the literature on career concerns. In the presence of asymmetric information, principals may wish to draw up an explicit incentive contract to resolve the moral hazard problem. However, as Fama (1980) argued, explicit contracts may not be required because career concerns induce agents to exert effort, i.e., career concerns

[3]The relevance of promotions as a sorting instrument is also shown on a more macro scale by Rosen (1982). His competitive labor market model illustrates how the use of promotions can explain the skewed distributions of firm size and earnings relative to the distribution of abilities.

provide an implicit incentive contract. Fama's idea initiated a great number of theoretical and empirical studies confirming his conjecture (see, e.g., Harris and Holmström, 1982; Chevalier and Ellison, 1999; Holmström, 1999).[4]

A first conceptualization of tracking in the sense of a tournament was made by Rosenbaum (1976).[5] His case study about the tracking system of an American high school revealed that students who dropped out of the college track had no chance of getting back into that track and very little chances of attending college. Thus, student performance prior to the track decision had a profound and enduring effect on later outcomes, providing students with a strong incentive to get and to stay in the college track.

Exploring the effect of different standards in high school testing, Becker and Rosen (1992) show that increasing standards and competition among students stimulates student effort and improves student achievement. Their model further shows that it may be preferable to separate students by ability instead of striving for a single nationally set standard. The reason is that a single grading scheme cannot give incentives for all students to provide a high effort level. In fact, their model predicts that mixing students of different abilities in one classroom reduces the effort of all students. High standards may additionally increase inequality because only high ability students may be able to cope with a high standard. Betts (1998) however disproves this hypothesis. According to his model, higher standards may create a pooling equilibrium, thereby increasing the earnings of both high ability and low ability workers.

A theoretical model for the incentive effects of tracking is provided by Eisenkopf (2009). He investigates the impact of performance-based selection at the transition to secondary school and comes to similar conclusions as Becker and Rosen (1992). In line with human capital theory, students will make an effort to improve their skills because higher skills are rewarded by the labor market in terms of higher wages. On top of that, tracking will induce particularly high effort levels in elementary school because employers use the attended secondary school track as a signal for ability and offer wages accordingly. To get admitted to the highest track, students in a tracking system will therefore exert more effort in elementary school than students in a comprehensive system.[6]

When investigating the relation between incentives and effort, the concepts of extrinsic and intrinsic motivation must be considered. A person who is motivated by the interest

[4]Note that risk-aversion and discounting limit the market's ability to incentivize agents (see Holmström, 1999).

[5]Rosenbaum formalized his tournament concept applied to the hierarchical structure of firms (see Rosenbaum, 1979; 1984).

[6]The incentives turn in secondary school: Students in a tracking system will exert less effort than students in a comprehensive system because their attended track already signals future employers their ability to some extent. In contrast, the strongest signal of students in a comprehensive is their GPA at the end of secondary school.

or the enjoyment in a task itself can be classified as intrinsically motivated, whereas a person who performs an activity to attain a certain outcome can be classified as extrinsically motivated (see Ryan and Deci, 2000). According to Ryan and Deci's Self-Determination Theory, motivation is not a global and undifferentiated concept that is synonymous with effort but rather a multidimensional concept that varies in terms of quality. Thus, motivation is of high quality when primarily based on intrinsic, integrated and identified regulations, and is of poor quality when primarily based on external and introjected regulations. Similarly, the Motivation Crowding Theory by Frey and Jegen (2001) suggests that external interventions may crowd-out intrinsic motivation. Applied to elementary school students and the transition phase between elementary and secondary school, marginal students may display lower levels of joy of learning than non-marginal students, since their parents are likely to push them relatively harder to make an effort to eventually get admitted to the higher track. In line with these theoretical considerations, the literature review on motivation and learning outcomes by Guay et al. (2008) concludes that "the more students endorse autonomous forms of motivation, the higher their grades are, the more they persist, the better they learn, and the more they are satisfied and experience positive emotions at school" (see Guay et al., 2008, p. 237).

Hypotheses

According to these theoretical considerations, I formulate the following hypotheses:

> **Hypothesis 1, a):** In a system with strict tracking based on test scores, marginal students increase their efforts relatively more than non-marginal students in the school term before the track decision is made.

The model by Eisenkopf (2009) predicts that tracking, based on academic performance, induces students to exert more effort in elementary school because the attendance of a higher secondary school track is associated with significantly higher career prospects. This prediction is supported by a large literature on career concerns. Since tracking in Bavaria is solely based on academic performance in elementary school, Bavarian students who are at the margin to be admitted to a higher track in the year before the track decision is made, should increase their efforts in the final year of elementary school relatively more than students who are clearly above the margin.

> **Hypothesis 1, b):** In a system with strict tracking based on test scores, marginal students decrease their efforts relatively more than non-marginal students once the track decision is made.

The literature on motivation and learning outcomes shows that students learn and perform the better, the higher their intrinsic motivation (Guay et al., 2008). In the year before

the track decision is made, marginal students in Bavaria, however, can be expected to experience relatively more extrinsic than intrinsic motivation, as most parents of marginal students will push their child to get into the highest track. Thus, the efforts of marginal students should decrease more than the efforts of non-marginal students once the track decision is made, and the extrinsic motivation strongly reduced.

> **Hypothesis 1, c):** The incentive effect of tracking on effort strongly depends on the existence of a strict admission policy for attending a certain track.

In Bavaria, the admittance to the secondary school tracks is solely based on the GPA at the end of the first half-year of fourth grade. Thus, marginal Bavarian students can be assumed to be highly (extrinsically) motivated to increase their effort in the school term before the track decision is made. In contrast, the admittance to a secondary school track in Hesse is independent of a student's GPA in elementary school. Thus, marginal Bavarian students should increase their effort much more than marginal Hessian students in the school term before the track decision is made.

> **Hypothesis 2, a):** In a system with strict tracking based on test scores, the increased effort of marginal students is accompanied by a decline in joy of learning compared to non-marginal students.

Guay et al. (2008) conclude that students are the more satisfied and experience the more positive emotions at school, the more they endorse autonomous forms of motivation. As marginal Bavarian students can be assumed to be more extrinsically motivated than non-marginal Bavarian students, I expect to find a relative decline in joy of learning among the marginal Bavarian students.

> **Hypothesis 2, b):** In a system with strict tracking based on test scores, marginal students improve on standardized competence measures relatively more than non-marginal students in the final year of elementary school.

To get admitted to the highest possible track, marginal students in Bavaria should increase their efforts in the school term before the track decision is made relatively more than non-marginal students. The relatively higher efforts should result in a relatively stronger improvement on standardized competence measures.

1.3 Institutional Setup

I investigate the incentive effects of tracking on effort and competence development within the framework of the German education system which I describe briefly in this section.[7]

[7] For a more detailed description, see for example Soskice (1994), or Winkelmann (1996).

The responsibility for the education system in Germany lies primarily with the states, though some generalizations are possible. Education starts with optional kindergarten, which is available to all children between three and six years of age. General compulsory schooling begins in the year in which a child turns six and involves a minimum of nine years of full-time schooling. In most states, students attend elementary school for four years, where they are taught the basics in reading, writing, and arithmetic (see Baumert et al., 2010, p. 95).[8] Additionally, the curriculum typically comprises local history classes as well as art, religious, and physical education.

After elementary school, students are tracked into one of three secondary school tracks. Basic and intermediate tracks (Hauptschule and Realschule) include schooling up to ninth and tenth grade, respectively, usually followed by vocational training, or the attendance of a higher secondary vocational school. The highest track (Gymnasium) qualifies for university studies and includes schooling up to 12th grade.[9] A fourth option are comprehensive schools: While 'cooperative' comprehensive schools provide all three school tracks in one school, 'integrative' comprehensive schools teach students of all aptitude levels together but offer university preparatory classes for the students who are doing well, general education classes for average students, and remedial courses for those who are struggling with the curriculum. In 2007, the year in which the students in the data set proceeded to secondary school, about 37 (44) percent of Bavarian (Hessian) fourth-graders enrolled in a Gymnasium, about 23 (16) percent in a Realschule, and about 39 (19) percent in a Hauptschule. Most of the remaining Hessian students (about 16 percent) enrolled in comprehensive schools (see Table D1-1A, Autorengruppe Bildungsberichterstattung, 2008, p. 253).[10]

The ultimate track decision is either made by the school, or the parents. If it is made by the school, the main criteria is the average grade obtained in math, German, and local history after the first half of fourth grade. According to the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK), no consideration should be given to parents' income, social class, or migration background.

In Bavaria, students with a grade point average (GPA) of 2.33 or better are allowed to attend the highest secondary school track (see Baumert et al., 2010, p. 99).[11] On trial, also students with a GPA worse than 2.33 may visit a Gymnasium. During a three-day

[8] As mentioned above, elementary school comprises six years in Berlin and Brandenburg.

[9] During the last decade, most states in Germany have reduced the duration of academic high school education from nine to eight years, but left the curriculum and the number of instruction hours up to the time of graduation unchanged.

[10] The finally remaining students in Bavaria and Hesse were enrolled in special types of secondary schools.

[11] In Germany, grades range from 1 to 5 with 1 being the best grade.

trial lesson in the receiving school, students' verbal and written competences in German and math are tested. If a student obtains an average grade of at least 3.5, he is allowed to attend the Gymnasium. The success rate among students attending trial lessons is at around 50 percent. However, only about four percent of the fifth-graders in Bavaria participate in a trial lesson (see ISB Staatsinstitut für Schulqualität und Bildungsforschung, 2009, p. 85 f.). Thus, the schools' track decisions can be considered as very strict.

In contrast, the ultimate track decision in Hesse is made by the parents. The class teacher gives parents a track recommendation, but this recommendation is not binding. Thus, no externally set standard is crucial for a student's secondary school track choice. Instead, the federal state law says that parents can claim a place for their child in the secondary school track they prefer (see Hessian education act, articles 70 to 77). It is then up to the regional education authorities to allocate the students to the schools within the chosen track.

The GPA distributions in our data among Bavarian and Hessian students who eventually attend the highest track are still similar, considering the GPAs at the end of third grade.[12] 89 percent of the Bavarian high school students had a GPA of 2.33 or better, and 98 percent a GPA of 2.66 or better. Among the Hessian students, 90 percent had a GPA of 2.33 or better, and 95 percent a GPA of 2.66 or better. The most striking difference between the Bavarian and the Hessian sub-sample, however, is the lowest GPA value found at the end of third grade among the students who eventually attended a Gymnasium: While this value is at 3.33 among the Bavarian students, the lowest GPA among the Hessian students is at 4.33, indicating the institutional difference between the two states.

Once the track decision is made, switches between secondary school tracks occur very rarely, and mainly downwards. At the time when the students in the data set proceeded to secondary school, only 2.6 percent of all students in Germany switched school tracks between seventh and ninth grade (see Tab. D1-4A, Autorengruppe Bildungsberichterstattung, 2008, p. 255). 65.6 percent of these track changes were from a higher to a lower track. After graduation from secondary school, there is the possibility to obtain a higher educational qualification through second- or third-chance education. However, only 4.4 percent of the first year university or university of applied sciences students obtained their university entrance certificate through second- or third-chance education in 2008 (see Table F1-4A, Autorengruppe Bildungsberichterstattung, 2010, p. 291). These figures show how crucial the secondary school track decision is for a student's future school career.

[12] The data provide no information on the GPA in fourth grade. Therefore, I present the GPA distribution in third grade as a proxy for the GPA distribution in fourth grade. The GPA at the end of third grade is furthermore very relevant as I use it to determine marginal and non-marginal students.

1.4 Data

To investigate whether tracking works as an incentive scheme, I use data of the BiKS-8-14 study, conducted by a team of researchers of the University of Bamberg. The goals of the study are to determine conditions for promoting children's linguistic and cognitive development as well as to explore the decision-making process concerning school enrollment and later on secondary school choice (see Mudiappa and Artelt, 2014).

Since autumn 2005, 2395 children from 155 different classes in 82 different schools in Bavaria and Hesse have been part of BiKS-8-14. About two thirds of the children attended a school in Bavaria, about one third a school in Hesse. Children were in third grade at the beginning of the BiKS study. The majority of the participating children moved on to a secondary school in summer 2007. In addition to the regular survey of competencies and characteristics of the children, including information on grades obtained in third grade, also parents and class teachers were interviewed, providing a wide range of background information at the family and school level.

I use data of the first three waves. The data of the first wave was collected at the end of third grade, the data of the second wave at the end of the first half of fourth grade, and the data of the third wave at the end of fourth grade. A student is defined as a marginal student if his GPA in math, German, and local history at the end of third grade is at 2.0 to 2.67. These students are exactly at or closely around the official threshold for high school entrance in Bavaria which is 2.33 at the end of fourth grade. The empirical threshold in the data for high school entrance in Hesse is also 2.33. I consider students with a GPA of 1.67 or better at the end of third grade as non-marginal students, serving as the control group. The final sample comprises a total of 1225 children (806 from Bavaria), of which 871 belong to the treatment group (567 from Bavaria) and 354 to the control group (239 from Bavaria). Summary statistics for selected variables are presented in Table 1.1. The first column shows statistics for the full sample, the second and the third column for the treatment and the control group, respectively. Both in the full and the split samples, the fraction of students attending a school in Bavaria is at around 65 percent, while the student-to-teacher ratio is at around 16.5. The fraction of male students in all three samples is at around 50 percent. 86 percent of the students speak German at home, whereas the fraction is slightly lower in the treatment group than in the control group (84 vs. 88 percent). In terms of academic achievement, the two groups differ significantly: The average grade in the treatment group is in each of the three subjects (math, German, and local history) about 0.8 grading steps lower than the average grade of the control group. Similarly, the fraction of parents with an upper secondary school degree as well as the highest ISEI level in the household is significantly lower in the treatment group than in the control group. Parents of marginal students are also less often married, and

Table 1.1: Descriptive Statistics

	(1) Full Sample	(2) Non-Marginal	(3) Marginal	(4) Δ in Means
Attends school in Bavaria	0.655	0.681	0.645	-0.0358
	(0.475)	(0.467)	(0.479)	(0.0281)
Student-to-teacher ratio	16.639	16.700	16.613	-0.0872
	(4.091)	(4.136)	(4.074)	(0.242)
Male	0.501	0.502	0.501	-0.00145
	(0.500)	(0.501)	(0.500)	(0.0296)
German at home	0.864	0.882	0.857	-0.0252
	(0.343)	(0.323)	(0.350)	(0.0207)
Parents married	0.827	0.864	0.812	-0.0523*
	(0.378)	(0.343)	(0.391)	(0.0228)
Parents immigrated	0.228	0.186	0.246	0.0603*
	(0.420)	(0.389)	(0.431)	(0.0248)
Upper secondary	0.486	0.658	0.415	-0.243***
	(0.500)	(0.475)	(0.493)	(0.0288)
Highest ISEI in household	53.586	58.724	51.418	-7.306***
	(15.958)	(15.842)	(15.513)	(0.946)
Employed	0.718	0.728	0.714	-0.0141
	(0.450)	(0.445)	(0.452)	(0.0272)
Partner employed	0.937	0.962	0.926	-0.0353*
	(0.243)	(0.193)	(0.262)	(0.0152)
German grade (3rd grade)	2.122	1.542	2.361	0.819***
	(0.629)	(0.499)	(0.509)	(0.0299)
Math grade (3rd grade)	2.035	1.433	2.283	0.849***
	(0.643)	(0.511)	(0.516)	(0.0304)
Local history grade (3rd grade)	1.939	1.302	2.202	0.900***
	(0.624)	(0.465)	(0.474)	(0.0279)
Observations	1225	354	871	1225

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

Note: Based on students from Bavaria with a GPA of 2.67 or better at the end of third grade (first wave). Students with a GPA of 2.0 to 2.67 are defined as marginal students. *Parents immigrated* is equal to one if at least one parent in the household immigrated. *Upper secondary* is equal to one if at least one parent in the household has an upper secondary school degree.

immigrated more often than parents of control students. In terms of employment, the differences between the groups are smaller and less significant.

To test the hypothesis that marginal students increase their efforts more than non-marginal students prior to the track decision, I make use of three items in the BiKS study that measure the willingness to exert effort. Two out of the three items were given both to students, teachers, and parents. The third item was only included in the student questionnaire. The three items are:

Table 1.2: Cronbachs Alpha

	Effort			Joy of Learning		
	Self-rating	Parents	Teachers	Self-rating	Parents	Teachers
1st wave	0.417	0.540	0.864	0.549	0.658	0.844
2nd wave	0.465	0.516	0.852	0.628	0.692	0.842
3rd wave	0.496	0.586	0.882	0.673	0.736	0.855

Note: Coefficients report the internal consistency estimate of reliability of the three effort and joy of learning measures within one wave (there are only two effort and joy of learning measures for parents, and only two effort measures for teachers).

1. I (He) give(s) up quickly when it becomes difficult.
2. I (He) make(s) an effort to solve difficult tasks.
3. I do not like to make an effort when it comes to learning.

The first three columns of Table 1.2 show coefficients of the internal consistency estimate of reliability (Cronbach's alpha) for the three effort measures for students, parents, and teachers. The theoretical value of alpha can take on values between zero and one, where higher values of alpha indicate a higher reliability. The teacher-assessments are most reliable, ranging from 0.85 to 0.88 in waves one to three, followed by the parental-assessments, ranging from 0.52 to 0.59, and the students' self-assessments, ranging from 0.42 to 0.50. These figures are comparable to the ones of Carbonaro (2005) and much higher than the alpha reported by Smerdon (1999), two studies from the education literature that used similar effort measures.

The scale of the items ranges from one to five in the teacher and the parental questionnaire, and from one to four in the student questionnaire. As the meaning of a high value (four or five) differs across the items, I first recode the scale of the second item such that a 'one' expresses a high effort level, and a 'five' a low effort level. Afterwards, I create a new variable containing the mean of the two (three) items. The eventual effort measures I use are standardized by wave to have a mean of zero and a standard deviation of one.[13]

The tournament-like situation at the end of elementary school should lead to higher effort levels of marginal students, in particular students who are at the margin to high school, as the high school track is associated with the highest career prospects. Higher effort levels may in turn lead to improved school performance and improved cognitive skills. The BiKS data allows to investigate both the effort provision, the school performance, and

[13]I more objective measure of effort might be the grades obtained in fourth grade. However, this information is not provided in the data set.

the development of cognitive skills of marginal and non-marginal students. In addition to the students' school performance and willingness to exert effort, the study also assesses the development of students' cognitive skills over time. The standardized tests comprise a reading competency test (ELFE 1-6), a mathematical competency test (DEMAT 3+ and DEMAT 4), an abstract reasoning test (CFT 20-R), a spelling test (DRT 3 and DRT 4), and a vocabulary test (CFT 20, WS). I use these test scores – standardized by wave to have a mean of zero and a standard deviation of one – to investigate whether the cognitive skills of marginal students improve more on average from third to fourth grade than the cognitive skills of non-marginal students.

If a child is at the margin to a higher track at the end of third grade, parents are likely to push the child to make an extra effort to eventually get into the highest track. Thus, marginal students should on average experience more extrinsic and less intrinsic motivation than non-marginal students. The BiKS study includes three items measuring the degree of joy of learning which I use to quantify students' intrinsic motivation. Two out of the three items were given to both students, teachers, and parents; the third item was only given to students and teachers. The three items are:

1. I (He) like(s) to go to school.
2. I (He) enjoy(s) learning in school.
3. I (He) actively participate(s) in class.

Columns four to six of Table 1.2 show the internal consistency estimates for the three joy of learning measures for students, parents, and teachers. The teacher-assessments are most reliable, ranging from 0.84 to 0.86 in waves one to three, followed by the parental-assessments, ranging from 0.66 to 0.74, and the students' self-assessments, ranging from 0.55 to 0.67.

To create the joy of learning measures, I proceed in the same way as with the effort measures: First, I create a new variable containing the mean of the two (three) items. Then, I standardize the measures by wave to have a mean of zero and a standard deviation of one.

1.5 Empirical Strategy

Uncovering the incentive effect of tracking is difficult primarily because of unobserved heterogeneity across students and classes. One of the most prominent factors that is usually not observed is the ability of students. Students with higher abilities may, for example, enjoy learning more than lower ability students, and therefore provide a higher

effort level. Students may also hold different beliefs about their own chances of academic success. Students who believe to succeed and expect to succeed in school will provide more effort because they anticipate that there will be a distinct "payoff" to their efforts. The effect of tracking on the provision of effort could then not be distinguished from the effect of certain individual characteristics. Another prominent factor that may lead to biased estimates is the peer group effect. On the one hand, there may be reverse causality when estimating the effect of a peer group on individual outcomes. On the other hand, students are likely to non-randomly select themselves into peer groups. For example, parents of high-achieving students may put their children into high-achieving classes.

Neglecting such unobserved heterogeneity across students and classes leads to results that cannot disentangle the effect of tracking from the influence of other unobserved factors. Controlling for family background and class characteristics certainly mitigates potential biases, but Table 1.1 raises doubts that the available variables can fully account for the nonrandom selection of students into treatment and control group as factors such as ability and peers are not directly captured.

Taking advantage of panel data, it is possible to control for unobserved but fixed heterogeneity across units of observation. In difference-in-differences (DiD) models, for example, only changes over time between units of observation are used to identify the effect of interest. The BiKS data allow to estimate such DiD models using the ordinary least squares (OLS) method. Equation (1.1) provides a linear specification:

$$Y_{ic}^{t} = \alpha_0 + \beta_0 \, Marginal_i^t + \beta_1 \, Wave_c + \gamma \, Wave_c \times Marginal_i^t \; + \; \delta \, X_i \; + \; \eta_{ic}^t, \quad (1.1)$$

where Y_{ic}^{t} denotes the outcome of interest of student i in wave c of type t, i.e., marginal or non-marginal. $Marginal_i^t$ is a dummy variable equal to one if a student is defined as a marginal student. $Wave_c$ captures wave-specific effects. X_i is a vector of demographic covariates comprising age, gender, parental education level, ISEI level, migration background, and language at home. Error terms η_{ic}^t are clustered at the teacher level to account for the presence of heteroscedasticity in the teacher-assessments. To obtain the estimates for the relative change in the outcomes between the first and the second, and the second and the third wave, I run two separate regressions only using observations of the respective waves.

The coefficient of interest is γ which measures the effect of being at the margin to the high school track on the change in the outcome variable between the first and the second, and the second and the third wave. Thus, hypothesis 1 a) will be fulfilled if γ is significantly positive, whereas hypothesis 1 b) will be fulfilled if γ is significantly negative. The key identifying assumption is that no other factors except for the status of being at the margin to a higher track at the end of third grade affected the outcomes of marginal and non-marginal students in fourth grade differently. Thus, the assumption is that the

underlying trends in the outcome variables would have been the same for both marginal and non-marginal students in the absence of tracking.

To investigate hypothesis 1 c), i.e., the importance of a strict admission policy for the incentive effect of tracking, I additionally include the dummy variable $Bavaria_i$ that is equal to one if a student attends a school in Bavaria, and zero otherwise, and interact it with the dependent variables of Equation (1.1). The resulting difference-in-difference-in-differences (DDD) model can be represented by Equation (1.2):

$$\begin{aligned} Y_{ic}^t = {} & \alpha_0 + \beta_0\, Marginal_i^t + \beta_1\, Wave_c + \beta_2\, Wave_c \times Marginal_i^t \\ & + \gamma_0\, Bavaria_i + \gamma_1\, Wave_c \times Bavaria_i + \gamma_2\, Marginal_i^t \times Bavaria_i \\ & + \gamma_3\, Wave_c \times Marginal_i^t \times Bavaria_i + \delta\, X_i + \eta_{ic}^t. \end{aligned} \tag{1.2}$$

The coefficient of interest is γ_3 which has the same interpretation as the coefficient γ in Equation (1.1). However, the DDD model controls for two further potentially confounding trends: First, it accounts for changes in the outcomes of marginal students across states, and second, for changes in the outcomes of both marginal and non-marginal Bavarian students.

As the BiKS study does not survey students before the end of third grade, it is not possible to show the trends in the outcome variables before the treatment, i.e., suggestive evidence for the validity of the common trend assumption. The robustness checks presented in the Section 1.6.1, however, suggest that the reported effects can be attributed to the status of being at the margin to a higher track.

The analysis may still provide misleading estimates because of an omitted variable bias arising from unobserved and uncontrolled differences between marginal and non-marginal students. If families in Bavaria (or Hesse) move – because of the tracking system – to another state where the secondary school track choice is independent of the teacher recommendation, the estimates would be biased. However, according to official statistics of the Federal Statistical Office (Destatis), Bavaria was the state with the highest net immigration in 2008, suggesting that a bias because of out-migration because of the elementary school system is unlikely (see Federal Statistical Office, 2010, p. 322f.). Similarly, statistics of the KMK and Destatis show that less than 0.2 percent of the Hessian students had been enrolled in one of the neighboring states in the school year 2014/15.

1.6 Results

In the following, I present estimates of the effect of being at the margin to a higher track on effort, joy of learning, and skill development. For the effort and joy of learning measures,

I focus on the teacher assessments since as are most reliable (see Table 1.2).

The section is divided into two parts. Section 1.6.1 presents estimates of hypotheses tests 1 a) to 1 c), i.e., the effect of being at the margin to the high school track on effort given the admission policy. Having presented the estimates of my baseline specification, I examine the sensitivity of the results to changes in sample restrictions and model specifications. Having demonstrated the robustness of the estimates, I further investigate whether the effects vary across subgroups. Afterwards, section 1.6.2 addresses the potential consequences of increasing effort on skill development and joy of learning, i.e., presents the results of hypotheses tests 2 a) and b).

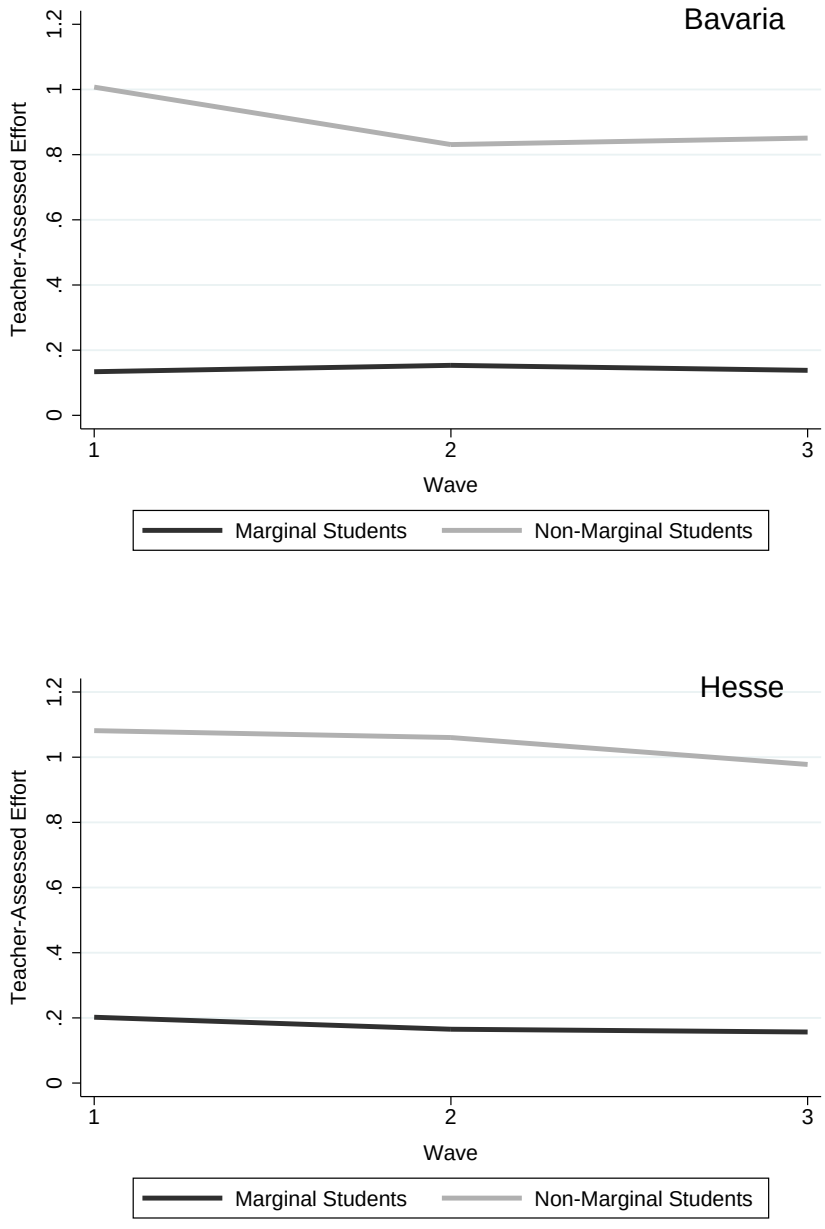
1.6.1 The Effect of Tracking on Effort

Figure 1.1 shows separately for Bavaria and Hesse the average teacher assessment of marginal and non-marginal students' effort provision over time, providing first evidence for the validity of hypotheses 1 a) to c). While in Bavaria, marginal students increase their efforts relative to non-marginal students in the school term before the track decision is made, i.e., between the first to the second wave, the effort provision of both marginal and non-marginal Hessian students decreases in the same period. After the track decision, between the second and the third wave, marginal students in Bavaria exert relatively less effort than non-marginal students. The reversed effect is found for the Hessian students.

The pattern shown in Figure 1.1 is confirmed by the DiD estimates of Equation (1.1): Columns one and two of Table 1.3 present estimates (with and without controls) for the effect of being at the margin to the high school track in Bavaria on marginal students' effort provision between the first to the second wave; columns three and four show estimates (with and without controls) of marginal students' effort provision between the second to the third wave. The coefficient of the interaction term *Wave x Marginal Student* represents γ, the parameter of interest in Equation (1.1). My baseline specification is the model including controls. The OLS regressions are based on the students with a GPA of 2.67 or better at the end of third grade. The control group (non-marginal students) consists of those students with a GPA of 1.67 or better. The inclusion of wave-specific controls does barely affect the coefficients, supporting the assumption that the estimation strategy accounts for unobserved but fixed differences between marginal and non-marginal students.

In line with hypothesis 1 a), I find a statistically and economically significant increase of about 18 percent of a standard deviation of marginal Bavarian students' effort provision relative to non-marginal Bavarian students' effort provision between the first and the second wave. Between the second and the third wave, i.e., after the track decision took place, the effort provision of marginal Bavarian students slightly decreases relative to the effort provision of non-marginal Bavarian students. However, this drop is not statistically

Figure 1.1: Effort Provision of Marginal and Non-Marginal Students Over Time



Note: Based on students from Bavaria and Hesse, respectively, with a GPA of 2.67 or better at the end of third grade (first wave). The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. Students with a GPA of 2.0 to 2.67 are defined as marginal students. The outcome variable is the teacher-assessment of students' effort provision standardized by wave and has a mean of zero and a standard deviation of one.

Table 1.3: The Effect of Tracking on Effort: Within Bavaria

	ΔWave 1 to 2		ΔWave 2 to 3	
	(1)	(2)	(3)	(4)
Wave x Marginal Student	0.196***	0.179***	-0.035	-0.037
	(0.058)	(0.063)	(0.052)	(0.056)
Marginal Student	-0.873***	-0.835***	-0.677***	-0.656***
	(0.060)	(0.066)	(0.064)	(0.076)
Wave Fixed Effect	-0.176***	-0.198	0.020	0.190
	(0.051)	(0.282)	(0.044)	(0.297)
Wave x Demographic Controls	No	Yes	No	Yes
Observations	1612	1540	1612	1540

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

Note: OLS regressions are based on students from Bavaria with a GPA of 2.67 or better at the end of third grade (first wave). The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. Students with a GPA of 2.0 to 2.67 are defined as marginal students. Students with a GPA better than 2.0 are defined as non-marginal students and represent the control group. The outcome variable is the teacher-assessment standardized by wave and has a mean of zero and a standard deviation of one. To obtain Δ *Wave 1 to 2* and Δ *Wave 2 to 3*, separate regressions were conducted. Wave-specific demographic controls include age, gender, parental education level, ISEI level, migration background, and language at home. Standard errors, reported in parentheses, are clustered at the teacher level.

significant, and hypothesis 1 b) consequently not strongly supported.

To test hypothesis 1 c), i.e., the importance of a strict admission policy for the incentive effect of tracking, I estimate Equation (1.2) using the sample of marginal and non-marginal students from both Bavaria and Hesse. The estimates, with and without wave-specific controls, are shown in Table 1.4. The coefficient of the interaction term *Wave x Marginal Student x Bavaria* represents γ_3, the parameter of interest in Equation (1.2). The model including controls represents my baseline specification.

If the strictness of the admission policy does not play an important role for the incentive effect of tracking, the Hessian students should behave similar to the Bavarian students, and the DDD coefficient should be zero. However, I find a statistically and economically significant increase of about 17 percent of a standard deviation of marginal Bavarian students' effort provision between the first and the second wave, which is almost identical to the DiD estimate of Equation (1.1). Thus, I find strong support for hypothesis 1 c). Although not statistically significant, the coefficient for the effect of being at the margin to the high school track on effort after the track decision took place, i.e., the change in effort between waves two and three, is more negative than the DiD estimate, providing some stronger support for hypothesis 1 b).

Table 1.4: The Effect of Tracking on Effort: Between Bavaria and Hesse

	ΔWave 1 to 2		ΔWave 2 to 3	
	(1)	(2)	(3)	(4)
Wave x Marginal Student x Bavaria	0.212**	0.172**	-0.109	-0.088
	(0.090)	(0.091)	(0.091)	(0.089)
Marginal Student x Bavaria	-0.074	0.001	0.218**	0.175
	(0.093)	(0.024)	(0.106)	(0.117)
Wave x Bavaria	-0.155**	-0.145**	0.102	0.111*
	(0.069)	(0.073)	(0.070)	(0.065)
Bavaria Fixed Effect	-0.074	-0.014	-0.229**	0.209*
	(0.093)	(0.022)	(0.099)	(0.109)
Wave x Marginal Student	-0.016	0.013	0.074	0.060
	(0.069)	(0.069)	(0.075)	(0.070)
Marginal Student	-0.879***	-0.187***	-0.895***	-0.826***
	(0.089)	(0.030)	(0.085)	(0.093)
Wave Fixed Effect	-0.021	-0.225	-0.083	-0.091
	(0.046)	(0.217)	(0.055)	(0.275)
Wave x Demographic Controls	No	Yes	No	Yes
Observations	2450	2316	2450	2316

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

Note: OLS regressions are based on students from Bavaria and Hesse with a GPA of 2.67 or better at the end of third grade (first wave). The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. Students with a GPA of 2.0 to 2.67 are defined as marginal students. Students with a GPA better than 2.0 are defined as non-marginal students and represent the control group. The outcome variable is the teacher-assessment standardized by wave and has a mean of zero and a standard deviation of one. To obtain the estimates of Δ *Wave 1 to 2* and Δ *Wave 2 to 3*, separate regressions were conducted. Wave-specific demographic controls include age, gender, parental education level, ISEI level, migration background, and language at home. Standard errors, reported in parentheses, are clustered at the teacher level.

Robustness Checks

In the following, I show that the results presented above are robust to a variety of alternative specification choices and validity checks. In particular, I perform a placebo test, and check whether the results differ when I vary the definition of being a marginal student.

If the effects are causal effects of being at the margin to the high school track, I should not find a change in effort within the group of students who are clearly above the margin at the end of third grade. The second column of Table 1.5 and 1.6, respectively, shows estimates of Equations (1.1) and (1.2) when defining students with a GPA of 1.67 at the end of third grade as marginal students, and students with a GPA better than 1.67 as non-marginal students. The results show that the so defined marginal and non-marginal

Table 1.5: Robustness Checks: Within Bavaria

	Baseline	Placebo Treatment	Without GPA of 2.0	Intermediate Track
ΔWave 1 to 2	0.179***	0.003	0.194***	0.164
	(0.063)	(0.102)	(0.076)	(0.106)
ΔWave 2 to 3	-0.037	0.127	-0.001	0.042
	(0.056)	(0.088)	(0.067)	(0.082)
Wave Fixed Effects	Yes	Yes	Yes	Yes
Group Fixed Effects	Yes	Yes	Yes	Yes
Wave x Demographic Controls	Yes	Yes	Yes	Yes
Observations	1612	458	1102	756

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

Note: OLS regressions are based on students from Bavaria. The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. The outcome variable is the teacher-assessment standardized by wave and has a mean of zero and a standard deviation of one. To obtain Δ*Wave 1 to 2* and Δ*Wave 2 to 3*, separate regressions were conducted. *Group Fixed Effects* refers to the fixed effect estimate of being a marginal student. Wave-specific demographic controls include age, gender, parental education level, ISEI level, migration background, and language at home. Standard errors, reported in parentheses, are clustered at the teacher level.

Table 1.6: Robustness Checks: Between Bavaria and Hesse

	Baseline	Placebo Treatment	Without GPA of 2.0	Intermediate Track
ΔWave 1 to 2	0.172*	-0.017	0.195*	0.220
	(0.091)	(0.140)	(0.109)	(0.140)
ΔWave 2 to 3	-0.088	0.106	-0.098	-0.106
	(0.089)	(0.133)	(0.104)	(0.152)
State Fixed Effects	Yes	Yes	Yes	Yes
Wave Fixed Effects	Yes	Yes	Yes	Yes
Group Fixed Effects	Yes	Yes	Yes	Yes
Wave x Demographic Controls	Yes	Yes	Yes	Yes
Observations	2316	680	1618	1090

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

Note: OLS regressions are based on students from Bavaria and Hesse. The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. The outcome variable is the teacher-assessment standardized by wave and has a mean of zero and a standard deviation of one. To obtain Δ*Wave 1 to 2* and Δ*Wave 2 to 3*, separate regressions were conducted. *Group Fixed Effects* refers to the fixed effect estimate of being a marginal student. Wave-specific demographic controls include age, gender, parental education level, ISEI level, migration background, and language at home. Standard errors, reported in parentheses, are clustered at the teacher level.

students do not change their efforts differently, neither between the first and the second, nor between the second and third wave. Thus, the placebo test supports a causal interpretation of the estimates reported in Tables 1.3 and 1.4.

In my baseline specifications, I consider students with a GPA of 2.0 to 2.67 at the end of third grade as marginal students, while the official threshold for attending high school in Bavaria is at 2.33. To check whether the results of my baseline specifications are solely driven by the better students within the group of marginal students – who potentially have a better parental background, or experience stronger support by their class teachers – I perform the baseline regressions, but exclude the students with a GPA of 2.0 at the end of third grade. The results presented in the third column of Table 1.5 and 1.6, respectively, are almost identical to my baseline results, showing that the effects on effort are not driven by the better marginal students.

Instead of considering the marginal high school students, I may also consider the marginal intermediate track students – who eventually have very similar incentives – to investigate the effects of tracking on effort. As in the main specification, the students with a GPA of 1.67 or better serve as the control group. The results presented in the fourth column of Table 1.5 and 1.6, respectively, support the estimates of the main specifications, as the effects of being at the margin to the intermediate track on effort are quantitatively and qualitatively very similar to the estimated effects of being at the margin to the high school track.

Subgroup Analysis

The analysis thus far has focused on the average effect of being at the margin to the high school track on effort, finding a significant and robust increase in effort of marginal Bavarian students in the school term before the track decision is made. Additionally, there could be important heterogeneity in the effect across subgroups. For example, many studies find that males and females behave and perform differently in competitive environments (see, e.g., Gneezy et al., 2003; Niederle and Vesterlund, 2007; Booth and Nolen, 2012), and a large literature establishes a link between parental background, competitive behavior, and educational achievement (see, e.g., Schnepf, 2003; Sirin, 2005; Almas et al., 2016). In the following, I investigate how the effects vary when I divide the sample into these subgroups. Due to the relatively small sample size of marginal Hessian students, and the resulting large standard errors, I perform the subsequent analysis by considering the Bavarian sample only.

The first two columns of Table 1.7 provide evidence for slightly heterogeneous effects with respect to gender. While both male and female marginal students significantly increase their effort prior to the track decision, boys reduce their effort afterwards, whereas

Table 1.7: Subgroup Analysis

	Gender		Parental Education		ISEI Level		Immigrant	
	Male	Female	HS	no HS	Top	Bottom	No	Yes
Δ Wave 1 to 2	0.165*	0.187**	0.165**	0.227**	0.094	0.330***	0.155**	0.300**
	(0.091)	(0.089)	(0.075)	(0.103)	(0.074)	(0.098)	(0.068)	0.115)
Δ Wave 2 to 3	-0.129	0.037	-0.066	0.006	0.002	-0.095	-0.038	0.007
	(0.081)	(0.073)	(0.070)	(0.082)	(0.067)	(0.082)	(0.058)	(0.132)
Wave Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Group Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Wave x Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	772	768	792	748	820	720	1322	218

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

Note: OLS regressions are based on students from Bavaria with a GPA of 2.67 or better at the end of third grade (first wave). The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. Students with a GPA of 2.0 to 2.67 are defined as students at-risk. Students with a GPA better than 2.0 are defined as non-marginal students and represent the control group. The outcome variable is the teacher-assessed effort provision standardized by wave and has a mean of zero and a standard deviation of one. *HS* is an abbreviation for high school degree, and is assigned to a student if at least one parent has a high school degree. To obtain Δ*Wave 1 to 2* and Δ*Wave 2 to 3*, separate regressions were conducted. *Group Fixed Effects* refers to the fixed effect estimate of being a marginal student. Wave-specific demographic controls include state dummies, age, gender, parental education level, ISEI level, migration background, and language at home. Standard errors, reported in parentheses, are clustered at the teacher level.

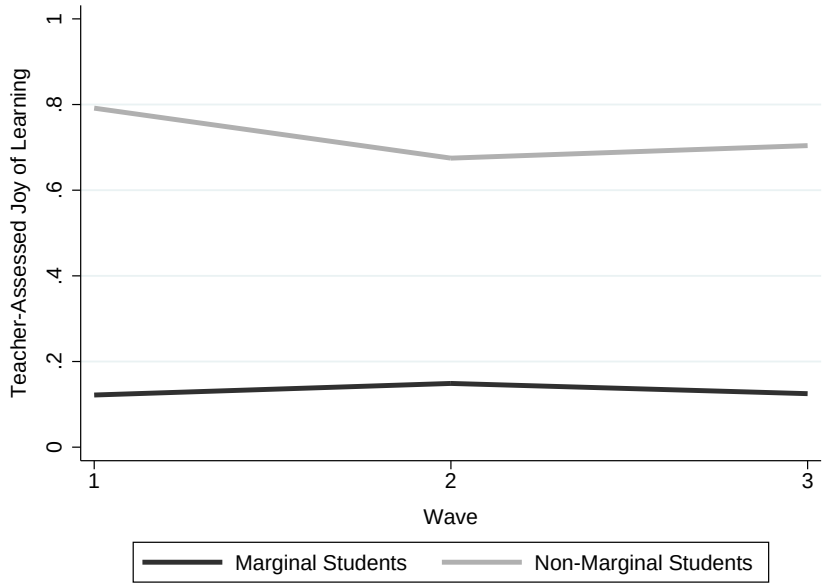
girls do not. This difference in the effort provision of boys and girls after the track decision is made is also statistically significant at the ten percent level.

Columns three and four of Table 1.7 show the estimates for the effect of being at the margin to the high school track on effort when dividing the sample by parents' highest secondary school degree. In column three, only students are considered who have at least one parent with a high school degree; column four comprises all students who have no parent with a high school degree. The results show that both groups significantly increase their effort in the school term before the track decision is made, while the coefficient for the group of students who have no parent with a high school degree is slightly larger. The difference between the coefficients, however, is not statistically significant. After the track decision is made, both groups do not change their efforts significantly.

Columns five and six of Table 1.7 show the results when splitting the sample at the median by the household highest ISEI ("International Socio-Economic Index") level, a widely used measure to classify a person's or a household's socio-economic status. The estimates suggest that the increase in effort prior to the admittance decision – which I find in the main specification – can to a large extent be explained by the students' socio-economic background: While the students coming from a household with an above-median socio-economic status do not significantly increase their effort between the first and the second wave, students coming from a household with a below-median socio-economic status increase their effort by about one third of a standard deviation. This difference in the effort provision between the two groups is also statistically significant at the ten percent level. After the track decision, both groups do not change their efforts significantly anymore.

The last two columns of Table 1.7 show the estimates when dividing the sample with respect to the migration background of students' parents. In column seven, only students are considered who have no parent with a migration background; in column eight, all students are considered who have at least one parent with a migration background. The results show that the students of both types significantly increase their efforts in the school term before the track decision is made, whereas the coefficient for the students with a migration background is about twice as large as the coefficient for the students with no migration background. The difference in the effort level between the two groups, however, is not statistically significant. After the track decision is made, marginal students with a migration background further increase their effort, although not statistically significantly, while native marginal students do not change their efforts.

Figure 1.2: Joy of Learning of Marginal and Non-Marginal Students Over Time



Note: OLS regressions are based on students from Bavaria with a GPA of 2.67 or better at the end of third grade (first wave). The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. Students with a GPA of 2.0 to 2.67 are defined as marginal students. The outcome variable is the teacher-assessment of students' joy of learning standardized by wave and has a mean of zero and a standard deviation of one.

1.6.2 The Effect on Joy of Learning and Skill Development

The BiKS data allow to investigate two further, highly relevant outcomes in the context of tournaments. First, a large literature provides evidence for a negative relation between extrinsic incentives and a person's intrinsic motivation (see e.g. Ryan and Deci, 2000; Frey and Jegen, 2001; Guay et al., 2008). Assuming that marginal students are pushed harder by their parents to make an effort than non-marginal students in the school term before the track decision is made, I expect to find a relative decline in marginal students' joy of learning. Second, the relatively higher effort of marginal students should result in a relatively stronger improvement in the standardized competence measures included in the BiKS data.

Figure 1.2 shows the development of the marginal and the non-marginal Bavarian students' joy of learning assessed by their class teachers. The trends are very similar to the trends found in Figure 1.1: While marginal students' joy of learning slightly increases in the school term before the track decision is made, non-marginal students display a significant drop. After the track decision, the trends reverse to some extent, and the difference in marginal and non-marginal students' joy of learning increases again.[14]

The DiD estimates of Equation (1.1) confirm this pattern: In the first column of Table 1.8, the estimates for the effect of being at the margin to the high school track in Bavaria

[14]The marginal and non-marginal Hessian students do not display a varying trend in joy of learning.

Table 1.8: Joy of Learning

	(1)	(2)
ΔWave 1 to 2	0.144***	0.162***
	(0.054)	(0.057)
ΔWave 2 to 3	-0.053	-0.041
	(0.045)	(0.045)
Wave Fixed Effects	Yes	Yes
Group Fixed Effects	Yes	Yes
Wave x Demographic Controls	No	Yes
Observations	1599	1529

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

Note: OLS regressions are based on students from Bavaria with a GPA of 2.67 or better at the end of third grade (first wave). The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. Students with a GPA of 2.0 to 2.67 are defined as marginal students. Students with a GPA better than 2.0 are defined as non-marginal students and represent the control group. The outcome variable is the teacher-assessed joy of learning measure standardized by wave and has a mean of zero and a standard deviation of one. To obtain Δ*Wave 1 to 2* and Δ*Wave 2 to 3*, separate regressions were conducted. *Group Fixed Effects* refers to the fixed effect estimate of being a marginal student. Wave-specific demographic controls include age, gender, parental education level, ISEI level, migration background, and language at home. Standard errors, reported in parentheses, are clustered at the teacher level.

Table 1.9: Competence Development

	Reading	Spelling	Vocabulary	Math	Logic
Δ Wave 1 to 2	0.170***		0.062	0.000	
	(0.064)		(0.045)	(0.084)	
Δ Wave 1 to 3	0.042	0.217***	0.091	0.153	-0.027
	(0.063)	(0.049)	(0.065)	(0.093)	(0.076)
Wave Fixed Effects	Yes	Yes	Yes	Yes	Yes
Group Fixed Effects	Yes	Yes	Yes	Yes	Yes
Wave x Demographic Controls	Yes	Yes	Yes	Yes	Yes
Observations	2239	1451	2242	2242	1452

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

Note: OLS regressions are based on students from Bavaria with a GPA of 2.67 or better at the end of third grade (first wave). The second wave was collected at the end of the first half of fourth grade, the third wave was collected at the end of fourth grade. Students with a GPA of 2.0 to 2.67 are defined as marginal students. Students with a GPA better than 2.0 are defined as non-marginal students and represent the control group. The outcome variables are standardized by wave and have a mean of zero and a standard deviation of one. *Group Fixed Effects* refers to the fixed effect estimate of being a marginal student. Wave-specific demographic controls include age, gender, parental education level, ISEI level, migration background, and language at home. Standard errors, reported in parentheses, are clustered at the teacher level.

on joy of learning without controls are shown; in the second column, the estimates with controls are shown. My baseline specification is the model including controls. In contrast to hypothesis 2 a), I find a significant increase by about 16 percent of a standard deviation of marginal students' joy of learning relative to non-marginal students' joy of learning between the first and the second wave. Between the second and the third wave, i.e., after the track decision, the effort provision of marginal students slightly decreases relative to the effort provision of non-marginal students. However, this drop is not statistically significant.

Table 1.9 presents the estimates of Equation (1.1) when using the standardized competence measures as the outcome variables. The results confirm hypothesis 2 b): Except for the logic test, the marginal students improve on the standardized competence measures relative to the non-marginal students between the first and the second, or the first and the third wave, respectively. In reading and spelling, these improvements are highly statistically significant, in the vocabulary and math tests they are almost statistically significant at the ten percent level. An explanation for the non-existent improvement in the logic test may be that logic is a competency that is not explicitly practiced in elementary school, and thus, rather reflects students' innate ability.

1.7 Discussion and Conclusion

In this paper, I provide evidence for an incentive effect of tracking on student outcomes measured prior to the track decision. Similar to rank-order tournaments in employment relationships, students who are at the margin to be admitted to a higher school track should increase their effort prior to the admittance decision as the attendance of a higher ability track is associated with significantly higher career prospects. Exploiting variation in the incentive intensity between marginal and non-marginal students in two German states over time, I find large and significant positive effects of being at the margin to a higher track on effort, motivation, and skill development. Strict admission policies for attending a certain track are crucial for the existence of the incentive effects of tracking.

My findings add to the findings by Koerselman (2013), the only study so far that investigated the incentive effects of tracking on student outcomes measured prior to the track decision. However, in contrast to Koerselman (2013), I observe student effort directly and not only indirectly via achievement test scores. Furthermore, my study does not suffer from the selection problems reported by Koerselman (2013).

The following conclusions can be drawn from my results: First, my results show that higher effort and student achievement may not only be achieved in secondary school through minimum competency tests or external exit exams, but also in elementary school through

tracking. This, secondly, shows that incentives from policies in earlier years of education must be taken into account when investigating the effect of policies at later ages. As individuals are forward-looking, measured outcomes are a result of policies at both earlier and later ages than the age of measurement. Third, there are methodological implications. For example, value added models using test scores to control for unobservables may provide biased estimates if the early age outcomes are affected by the policy under consideration. The existence of incentive effects may also invalidate the use of early outcomes in placebo tests. In a carefully controlled experiment, there is no difference in pre-treatment outcomes between the treatment and the control group. In natural experiments, however, subjects may be aware of their future treatment status and behave accordingly. For example, Pischke and Manning (2006) find that test score growth in the UK between age 7 and 11 is correlated with tracking policies after the age of 11.

Finally, it is not clear that tracking is a good measure to provide incentives in schools as it has a cost associated with it in terms of inequality and intergenerational mobility (see e.g. Hanushek and Wößmann, 2006; Malamud and Pop-Eleches, 2011). If comprehensive school students do indeed catch up with early tracking students in terms of achievement, as suggested by Hanushek and Wößmann (2006), the positive effects of early tracking on later age outcomes may not be very large. It is furthermore not clear whether increased test scores reflect improved learning or rather improved test-taking skills (see e.g. Jacob, 2005; Almlund et al., 2011). At least, my results do not provide evidence for a direct negative effect of tracking on intrinsic motivation as suggested by Jürges et al. (2012).

References

Almas, I., Cappelen, A. W., Salvanes, K. G., Sorensen, E. O., and Tungodden, B. (2016). Willingness to compete: Family matters. *Management Science*, 62(8):2149–2162.

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Chapter 1 - Personality psychology and economics. In Eric A. Hanushek, S. M. and Wößmann, L., editors, *Handbook of The Economics of Education*, volume 4 of *Handbook of the Economics of Education*, pages 1 – 181. Elsevier.

Amabile, T. (1996). *Creativity In Context: Update To The Social Psychology Of Creativity.* Westview Press.

Autorengruppe Bildungsberichterstattung, editor (2008). *Bildung in Deutschland 2008. Ein indikatorengesttzter Bericht mit einer Analyse zu bergngen im Anschluss an den Sekundarbereich I.* Bertelsmann, Bielefeld.

Autorengruppe Bildungsberichterstattung, editor (2010). *Bildung in Deutschland 2010. Ein indikatorengesttzter Bericht mit einer Analyse zu Perspektiven des Bildungswesens im demografischen Wandel.* Bertelsmann, Bielefeld.

Baumert, J., Maaz, K., Gresch, C., McElvany, N., Anders, Y., Jonkmann, K., Neumann, M., and Watermann, R. (2010). *Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten. Zusammenfassung der zentralen Befunde.* Bundesministerium für Bildung und Forschung, Referat Bildungsforschung, Bonn.

Becker, W. and Rosen, S. (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review*, 11(2):107–118.

Betts, J. R. (1998). The impact of educational standards on the level and distribution of earnings. *The American Economic Review*, 88(1):266–275.

Booth, A. and Nolen, P. (2012). Choosing to compete: How different are girls and boys? *Journal of Economic Behavior & Organization*, 81(2):542 – 555.

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education*, 78(1):27–49.

Chevalier, J. and Ellison, G. (1999). Career concerns of mutual fund managers. *The Quarterly Journal of Economics*, 114(2):389–432.

Dechenaux, E., Kovenock, D., and Sheremeta, R. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *The American Economic Review*, 101(5):1739–1774.

Eisenkopf, G. (2009). Student selection and incentives. *Zeitschrift für Betriebswirtschaft*, 79(5):563–577.

Eisenkopf, G. and Teyssier, S. (2013). Envy and loss aversion in tournaments. *Journal of Economic Psychology*, 34(C):240–255.

Fama, E. (1980). Agency problems and the theory of the firm. *Journal of Political Economy*, 88(2):288–307.

Federal Statistical Office, editor (2010). *Wirtschaft und Statistik*. Federal Statistical Office, Wiesbaden.

Frey, B. and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5):589–611.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review*, 87(4):746–755.

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

Guay, F., Ratelle, C. F., and Chanal, J. (2008). Optimal learning in optimal contexts: The role of self-determination in education. *Canadian Psychology*, 49(3):233.

Hanushek, E. A. and Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences- in-differences evidence across countries. *Economic Journal*, 116(510):C63–C76.

Harris, M. and Holmström, B. (1982). A theory of wage dynamics. *The Review of Economic Studies*, 49(3):315–333.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182.

ISB Staatsinstitut für Schulqualität und Bildungsforschung, editor (2009). *Bildungsbericht Bayern 2009*. Kastner AG, Wolnzach.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89(5-6):761–796.

Jürges, H., Schneider, K., Senkbeil, M., and Carstensen, C. H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31(1):56–65.

Koerselman, K. (2013). Incentives from curriculum tracking. *Economics of Education Review*, 32:140–150.

Lazear, E. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–64.

Malamud, O. and Pop-Eleches, C. (2011). School tracking and access to higher education among disadvantaged groups. *Journal of Public Economics*, 95(11-12):1538–1549. Special Issue: International Seminar for Public Economics on Normative Tax Theory.

Meghir, C. and Palme, M. (2005). Educational reform, ability, and family background. *The American Economic Review*, 95(1):414–424.

Mudiappa, M. and Artelt, C. (2014). *BiKS - Ergebnisse aus den Längsschnittstudien. Praxisrelevante Befunde aus dem Primar- und Sekundarschulbereich.* University of Bamberg Press, Bamberg.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Pischke, J.-S. and Manning, A. (2006). Comprehensive versus selective schooling in England in Wales: What do we know? Working Paper 12176, National Bureau of Economic Research.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63.

Rosen, S. (1982). Authority, control, and the distribution of earnings. *Bell Journal of Economics*, 13(2):311–323.

Rosenbaum, J. (1976). *Making inequality: The hidden curriculum of high school tracking.* Wiley-interscience publication. Wiley, New York.

Rosenbaum, J. (1984). *Career mobility in a corporate hierarchy.* Academic Press, Orlando, FL.

Rosenbaum, J. E. (1979). Tournament mobility: Career patterns in a corporation. *Administrative Science Quarterly*, 24(2):220–241.

Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, pages 68–78.

Schnepf, S. V. (2003). Inequalities in secondary school attendance in germany.

Sirin, S. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3):417–453.

Smerdon, B. (1999). Engagement and achievement: Differences between african-american and white high school students. *Research in Sociology of Education and Socialization*, 12:103–34.

Soskice, D. (1994). Reconciling markets and institutions: The german apprenticeship system. In *Training and the Private Sector: International Comparisons*, pages 25–60. National Bureau of Economic Research, Inc.

Winkelmann, R. (1996). Employment prospects and skill acquisition of apprenticeship-trained workers in germany. *Industrial and Labor Relations Review*, 49(4):658–672.

Chapter 2

The Effect of a Compressed High School Curriculum on University Performance

2.1 Introduction

The optimal design of the schooling system is a fundamental issue for economic policy. Concerning the optimal length of schooling, policy makers face a difficult trade-off. On the one hand, instruction time has been shown to be positively related to academic achievement (see, e.g., Bellei, 2009; Wößmann, 2003), while more years of schooling have been shown to yield sizable monetary and non-monetary benefits (Card, 1999; Lochner, 2011). On the other hand, the entry into the labor force is delayed and the duration of gainful employment reduced with an increasing length of schooling. Hanushek and Wößmann (2008) furthermore show that cognitive skills rather than mere school attainment determine economic well-being, and that the quality of school institutions is decisive. This raises the question of whether school resources are used efficiently, and whether the distribution of instruction time, and hence the length of schooling, is optimal. So far, most of the research designs are not able to investigate this question, as most school reforms simultaneously affect instruction time, curriculum covered, and school duration (Patall et al., 2010). A high school reform, recently implemented in Germany, provides a setting to investigate this issue.

Since the early 2000s, most German states have reduced the duration of academic high school education from nine to eight years, but left the curriculum and the number of instruction hours up to the time of graduation unchanged. As a consequence, the number of instruction hours per day, and hence the learning intensity increased. This is the most fundamental reform of the German education system in the last decades and presumably the most controversially debated one. The main concern is that the higher learning intensity may have negative consequences on children's development, in particular on their learning and human capital accumulation (see Lehn, 2010). Thus, opponents of the reform assume that the ratio of time spent in school to time for recreation was closer to optimal under the old system. However, evidence supporting this claim is missing.

This paper helps to fill this gap by providing first evidence of longer-term effects of a reduction in the duration of high school education. Using a unique data set of university students, we analyze the effect of the reform on cognitive skills. A major advantage of the German reform for analyzing the effect of a reduction in the duration of high school education is the way the reform was implemented: The German states introduced the eight-year system only in academic high schools, whereas it was not introduced in other high school types. This allows us to disentangle the effect of the reform from cohort, state, and school-type effects by estimating a difference-in-differences (DiD) model. Another major advantage of our data is that we observe treatment and control students taking the same exams. Consequently, their performance is highly comparable.

We use administrative, student-level panel data of the registrar's office of the University

of Konstanz, located in the state of Baden-Württemberg. Most students in Germany choose a university close to their home town. Thus, the majority of the students in our sample graduated from high school in Baden-Württemberg, representing our treatment state.[1] Baden-Württemberg introduced the reform in academic high schools in the school year 2004/05. The first students with an eight-year high school program (G8 students) enrolled at the University of Konstanz in fall 2012. At the same time, vocational high school students with a nine-year high school program (G9 students) enrolled. We exploit this variation across cohorts and between high school types for identification. The main advantage of using variation within a state is that estimates are not affected by unmeasured state-level policies, as long as all schools are equally affected (Hanushek et al., 1996).

The existing evidence on the short-run effects of the reform does not generate clear predictions about the reform's long-run effects. On the one hand, fifteen-year-olds have been shown to benefit from the new system in terms of PISA performance (Homuth, 2012; Andrietti, 2015). High school graduation rates have furthermore been shown to be unaffected, while students are on average almost one year younger at the time of graduation(Huebener and Marcus, 2017). On the other hand, grade repetition rates in grade ten increased under the new system (Huebener and Marcus, 2017), and math scores in the final high school exam decreased, at least in the first G8 cohort in Saxony-Anhalt (Büttner and Thomsen, 2013). Even if the latter result applies to all German states and all affected cohorts, the reform may still positively affect university performance if it improves at the same time students' non-cognitive skills, such as the ability to cope with stress.[2]

We find that students of the first G8 cohort who graduated from academic high schools together with students of the last G9 cohort (in the following also referred to as double cohort students) performed similar to students of the control group. Considering the second G8 cohort, we find significant positive effects on the average grade obtained and the likelihood to fail an exam. These positive effects stem particularly from female students. We find no significant effects for male students. Robustness checks support these findings. Given the one-year younger student body, our results suggest that the reform lowers the opportunity costs of education and facilitates an earlier labor market entry.

The paper proceeds as follows: In section 2.2, we present background information on the German education system, and especially the German high school reform. Section 2.3 gives an overview of the related literature. Section 2.4 describes our data. In section 2.5, we present our identification strategy, provide graphical support for its validity, and discuss potential threats. Section 2.6 presents our findings, and section 2.7 concludes.

[1] As a robustness exercise, we compare our findings with the treatment effect obtained for Bavarian G8 students.

[2] See, e.g., Carneiro et al. (2007) on the role of cognitive and non-cognitive skills for later outcomes.

2.2 Institutional Setup

The German education system is characterized by a distinct federalism, though some nationwide standards have been pursued as well. In 1955, the *Düsseldorfer Abkommen* (Düsseldorfer convention) set the years of schooling required to earn the *Abitur* (high school diploma) to 13 years. Following this convention, students in most West German states spent four years in primary school and nine years in secondary school until receiving the high school diploma (see Kühn et al., 2013).[3] In 1997, "The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany" (KMK), representing the most important interface of the German states within the national education policy, decided that the number of instruction hours up to the high school diploma must be identical across states, comprising 265 so-called "Jahreswochenstunden". The number of school years, however, was allowed to vary.

These criteria set in 1997 apply to both academic and vocational high schools. However, while both academic and vocational high schools provide general education and students can earn a high school diploma which qualifies for university studies at any German university, G8 programs were only implemented in academic high schools.[4] We exploit this variation between school types (and across cohorts) to identify the effect of a reduction in years of schooling on academic achievement at the tertiary level.

Between 2001 and 2007, 13 out of 16 states implemented the eight-year high school program. The only state still planning to keep the old program is Rhineland-Palatinate.[5] At the time of implementation, students were in fifth grade in most states.[6] As a consequence of the reform, the learning intensity increased for G8 students: While G9 students have on average only a bit more than 29 instruction hours per week, G8 students spend on average about 33 hours per week in school. This increase is even more pronounced in grades seven to ten, because the number of weekly instruction hours was basically left unchanged in grades five and six.

The policy makers' primary goal was to reduce the labor market entry age. Before the reform, the age of German high school graduates was high by international comparison: While the average graduation age in the OECD is at about 18 years, German high school

[3] After primary schooling, students are tracked into one of three secondary school tracks. The basic and intermediate tracks include schooling up to grade 9 and 10, respectively, usually followed by vocational training.

[4] Vocational high schools differ from academic high schools to the extent that vocational high school students additionally specialize in a work-related discipline in the last three to five years (the duration varies across states).

[5] Today some states discuss returning to G9 because of public concerns.

[6] With the following exceptions: In Saxony-Anhalt (ST) and Mecklenburg-Western Pomerania (MV) the reform was implemented in grades five to nine. In Bavaria (BY) and Lower Saxony (NI) it was implemented in grade five and six.

graduates were aged between 19 and 20 (see OECD, 2014). By reducing the graduation age, policy makers aimed to counter the effects of demographic change and to ensure sustainability of the social security systems by increasing the number of workers available to the labor market. At the same time, the quality of the high school diploma should not be reduced by the reform. Therefore, the curriculum and the number of instruction hours were remained unchanged, thus increasing the international competitiveness of German students (see Klemm, 2008).

2.3 Related Literature

There are only a few studies that identify the direct effect of a variation in the years of schooling on educational outcomes. Most research addressing school time variations estimate a mixed effect of differences in the amount and the distribution of instruction time, and the curriculum covered.[7]

Krashinsky (2014) and Morin (2013) explore a policy change in Canada similar to the German G8 reform that reduced high school by one year. However, in contrast to the G8 reform, only part of the curriculum remained unchanged, and the choice of the program was not random. Based on an instrumental variable strategy, Krashinsky (2014) finds significant negative effects of the shortened high school education on later educational outcomes. Morin (2013) explores the fact that the mathematics curriculum was shortened from five to four years while the length of the biology curriculum for the same students remained unchanged. His results point to a small positive effect of an extra year of high school mathematics on university performance.

A relatively clear setting for the analysis of the impact of learning intensity on academic achievement is provided by the introduction of a four-day-school week. In this case, a given amount of instruction time is distributed over fewer days. This increases the learning intensity during the school days, but also increases the number of recreation days. In this setting, Anderson and Walker (2015) identify negative effects for math and reading skills. Eren and Millimet (2007) explore the effects of the length of the school year, the

[7]This issue applies to most studies investigating cross-country variation, variations in weather-related closing days, and remediation programs. Cross-country or cross-state estimates are affected by all three channels. These studies mainly find a small positive or zero correlation between instruction time and educational outcomes (Wößmann, 2003; Lavy, 2015; Lee and Barro, 2001; Mandel and Süssmuth, 2011). Weather-related changes in school days result in a reduction of instruction time keeping the curriculum constant. These closing days have been found to have negative effects on academic achievement (Marcotte, 2007; Marcotte and Hemelt, 2008; Hansen, 2011). Studies exploring variations in the school system, such as changes from part-time to full-time schooling (Bellei, 2009) or the length of the school year (Eren and Millimet, 2007; Lavy, 2012), analyze a simultaneous change of the instruction time, the curriculum, the academic and non-academic activities and the financial budget. The reported effects are mainly positive. Positive effects have also been reported for remediation programs (see Battistin and Meroni (2016) for a detailed overview).

number of class periods per day, and the average length per class period in the United States. They find that shorter class periods, but more class per day, is associated with higher mean student achievement. This effect is more pronounced for students belonging to the lower quantile of the test score distribution.

Similar to the G8 reform, Pischke (2007) analyzes a change in the German school system taking place in the 1960s that led to a shorter primary school year while keeping the curriculum constant. Thus, Pischke (2007) also investigates the effect of a change in the length of schooling which is independent of the curriculum studied. He finds that the short-school years were associated with an increase in grade repetition in primary school, and a smaller number of students attending higher secondary school tracks. However, his results also show that the affected students were able to compensate for these short-run effects, as he finds no effect on later employment and wages.

In contrast to the reform analyzed by Pischke (2007), the G8 reform takes place at the secondary school level, and the content of the eliminated year is distributed over several years. First insights into the effects of this particular reform were gained from a survey conducted among double cohort graduates in the state of Saxony-Anhalt. Büttner and Thomsen (2013) find that the reform reduced the final examination scores in Mathematics for both genders, and in English for females. Meyer and Thomsen (2016) reveal that female G8 students delay university enrollment and are more likely to start a vocational training. Meyer and Thomsen (2013) show that there are almost no differences in motivation and perception of stress between G8 and G9 student at the university. In contrast, Quis and Reif (2017) find that particularly female G8 students experience a higher stress level and more mental health problems compared to G9 students in the double cohort in Baden-Württemberg. They find no health effects for male students.

The analysis of a double cohort has some shortcomings. Firstly, potential reform effects cannot be separated from general trends as only a single point in time is considered. Secondly, other than for later G8 cohorts, incentives and mental pressure may be different for G8 students of the double cohort as they directly compete with older students for limited resources (e.g., for university places). The double cohort of Saxony-Anhalt is moreover special to the extent that the reform was implemented in ninth grade, resulting in particularly high numbers of weekly instruction hours in grades nine to twelve. These G8 students were therefore differently affected by the reform than other G8 students.

Several more recent studies overcome these shortcomings. These confirm that G8 has an effect on skills while students are enrolled in high school, but there is no clear evidence that differences persist in the long-run. Homuth (2012) and Andrietti (2015) find significant positive effects on the reading, mathematics, and science literacy skills of fifteen-year-olds using the PISA data. These results are not surprising as G8 students

already received more instruction time at the time of PISA than the control students. Dahmann (2017) arrives at similar conclusions. She finds that fluid intelligence, i.e., the capacity to think logically and solve problems in novel situations, remained unaffected by the reform, and crystallized intelligence, i.e., the ability to use skills, knowledge, and experience, improved for male students, but only if G8 and G9 students are compared within the same grade. By the time of graduation G8 students seem to possess the same competences as G9 students.[8]

There is also no evidence that G8 affects high school graduation and university enrollment. G8 students only seem to be more likely to decelerate their educational career by repeating grades and taking a year off after high school graduation. Using aggregated administrative records Huebener and Marcus (2017) reveal that G8 reduces the average graduation age by about 10 months and increases the fraction of students repeating a grade by about three percentage points. The high school graduation rate is, however, not affected. Meyer et al. (2015) show that university enrollment in the first year after graduation is reduced by about 15 percentage points. Considering planned enrollment beyond the first year, the effect becomes much smaller and disappears in most specifications.

Our paper complements the literature on years of schooling, and in particular the literature on the G8 reform, by investigating the effects on academic achievement at the tertiary level, i.e., longer-term effects. Using within state variation, our estimates are not affected by unmeasured policies being in place at the state-level. General trends are moreover accounted for by investigating within-exam variation over time.

2.4 Data

We use unique, student-level panel data on university performance provided by the registrar's office of the University of Konstanz. Located in the southernmost part of the state of Baden-Württemberg, the majority of the students attended high school in Baden-Württemberg (about 76 percent). Our analysis focuses on the students who graduated from high school in Baden-Württemberg.

The data set contains information on university performance and program choice of all undergraduate students who graduated from high school between 2009 and 2013. In total, the data comprise about 8000 students, adding up to approximately 160.000 single student-exam observations. In addition to the grades, we have information about the students' major, exam dates, i.e., the exam semester and whether the exam was written at the first or at the second exam date, the number of preceding attempts, as well as the

[8]Based on SOEP data, Dahmann and Anger (2014) furthermore show that G8 students are more extroverted and less emotionally stable. Based on the double cohort in Saxonvy Anhalt, Thiel et al. (2014) find no significant effects on personality traits.

semester a student was enrolled in when she took the exam. Concerning the high school career, we have information on the students' overall grade point averages, the type of high school a student attended, and the place and date of issue of the high school diploma. Further, the data set contains information about each student's year and month of birth, gender, and nationality. We use this information to assign students to a G8 or a G9 cohort in the following way:

First, every student who did not graduate from an academic high school is identified as G9 student as the reform was only implemented at academic high schools. Second, we can identify every student who had graduated from an academic high school before the double cohort graduated as a G9 student. Conversely, every student who graduated from an academic high school after the double cohort can be identified as a G8 student. In Baden-Württemberg, double cohort students graduated from academic high schools in 2012. Consequently, we identify students who obtained their high school diploma in Baden-Württemberg up to 2011 as G9 students, and students who graduated from an academic high school in Baden-Württemberg after 2012 as G8 students.

Concerning the identification of the double cohort students, we make use of the fact that the cut off date for school enrollment for every student in our sample was the 30th of June. A child who turned six until this cut off date was enrolled in primary school in the same year.[9] Therefore, it is possible to define whether a double cohort student belongs to a G8 or G9 cohort based on his date of birth and his high school graduation year. Students who graduated from an academic high school in Baden-Württemberg in 2012 and were born before the 1st of July 1993 were not affected by the reform. In contrast, students who graduated from an academic high school in Baden-Württemberg in 2012 but were born on the 1st of July 1993 or later are identified as G8 students. To validate our assignment, we conducted a survey among all currently enrolled undergraduate students, asking them to state whether they belonged to a G8 or a G9 cohort. This information was inquired twice to minimize wrong statements, i.e., once by asking the student if he belonged to a G8 cohort, and once by asking if he belonged to a G9 cohort. In total, 1987 students replied, 406 of them stating that they belonged to a G8 cohort, and 1581 stating that they belonged to a G9 cohort. Each of these three figures represent roughly one third of the respective, currently enrolled student population. Thus, no group of students was under- or over-represented in our survey. When we compare the questionnaire data with our processed university data, we find that about four percent of the G8 students of our

[9]Since 1998, parents in Baden-Württemberg have the opportunity to pre- or postpone the school enrollment of their child by one year. However, only about 10 percent of the children in Baden-Württemberg were enrolled earlier or later than regularly between 2005 and 2013 with fractions being very stable (see Table 5.1 Statistisches Bundesamt, 2014, p. 268). We further address this potential issue as discussed below.

baseline sample who participated in the survey were incorrectly specified as G9 students by our procedure. Similarly, about two percent of the G9 students participating in the survey were incorrectly specified as G8 students.

These misassignments may occur because some students were enrolled earlier or later than required, skipped a grade, or were retained. To avoid such misassignments, we exclude students from our baseline samples whose actual years of schooling (calculated by their age at graduation minus six) do not match with their expected years of schooling. Consequently, we restrict our samples to 18- to 19-year-old G8 high school graduates, and to 19- to 20-year-old G9 high school graduates, i.e., to students with a regular high school career. The exclusion of students who were retained is furthermore necessary because retained primary school students may have graduated in 2012, eventually belonging to the first G8 cohort. Similarly, students who originally started in the first G8 cohort but were retained once during high school eventually graduated in 2013. Thus, the fraction of repeaters in the first G8 cohort is by construction lower. Another issue concerning the sample composition is the fact that many students start their studies not immediately after their high school graduation but several months or years later[10], and this time span could be related to a student's motivation and other unobserved characteristics. The G8 students of the first cohort, however, could take a break of at most two and a half years between their high school graduation and their university enrollment. We therefore restrict the sample for analyzing the effects on the first G8 cohort (in the following also referred to as 2012 sample) to students who enrolled at the University of Konstanz at most two and a half years after their high school graduation. For the same reason, we restrict the sample for analyzing the effects on the second G8 cohort (in the following also referred to as 2013 sample) to students who started their studies at most one and a half years after their high school graduation.

We make another restriction at the class level. In our baseline samples, we only consider academic achievement from the first two semesters, taken at the first attempt. Most of the classes taught within the first two semesters of a Bachelor's degree program are mandatory classes, and students are registered for the final exam automatically at the first attempt. Consequently, we rule out that our results are driven by a different class selection of G8 students. The final 2012 sample comprises 2562 students, of which 550 belong to the treatment group. The final 2013 sample consists of 2326 students, of which 505 belong to the treatment group. Figure 2.1 shows the number of academic and vocational high school students from Baden-Württemberg pooled by cohort. In 2012, the year in which the G8 and G9 cohorts graduated jointly, the number of academic high school students

[10]About 55 percent of the students start their studies within the year of their high school graduation, about 85 percent within one and a half years, and about 4 percent after more than two and a half years.

basically doubled. Considering the graph of the vocational high school students, no similar pattern shows up; the number of vocational high school students is basically constant over time. Furthermore, Figure 2.2 shows that the ratio of vocational to academic high school graduates in our sample is representative for the statewide ratio of vocational to academic high school graduates.[11] This is a first indicator that our baseline sample is not biased by a different selection into high school types due to the reform.

We study the following outcomes: grades obtained, the likelihood to drop out of university within the first two semesters, and the average time span between high school graduation and university enrollment. University grades range from 1 to 5 with 1 being the best grade and 5 being the lowest. To make grading between faculties comparable we standardize the grades by exam level to have a mean of zero and a standard deviation of one. We then investigate the effect of the reform on the average grade obtained, the likelihood to fail an exam and the likelihood to obtain a top grade. The latter achievement measures are binary variables that are equal to one if a student obtained a 5, which is equivalent to failing an exam, or a grade below 1.5, respectively. The outcome variable "Dropout" is equal to one if a student did not proceed to the third semester. "One-year break" is equal to one if a student enrolled earliest 14 month after high school graduation.[12] When we estimate the effect on grades, the regression models include dummies for sex, nationality, majors, semesters, and exams. The estimations on the average time span between high school graduation and university enrollment and the likelihood to drop out of university only include dummies for sex, nationality, and majors. When estimating the effect on the outcome variables "Dropout" and "One-year break", one further restriction is necessary: In both cases, each student must be considered only once.[13]

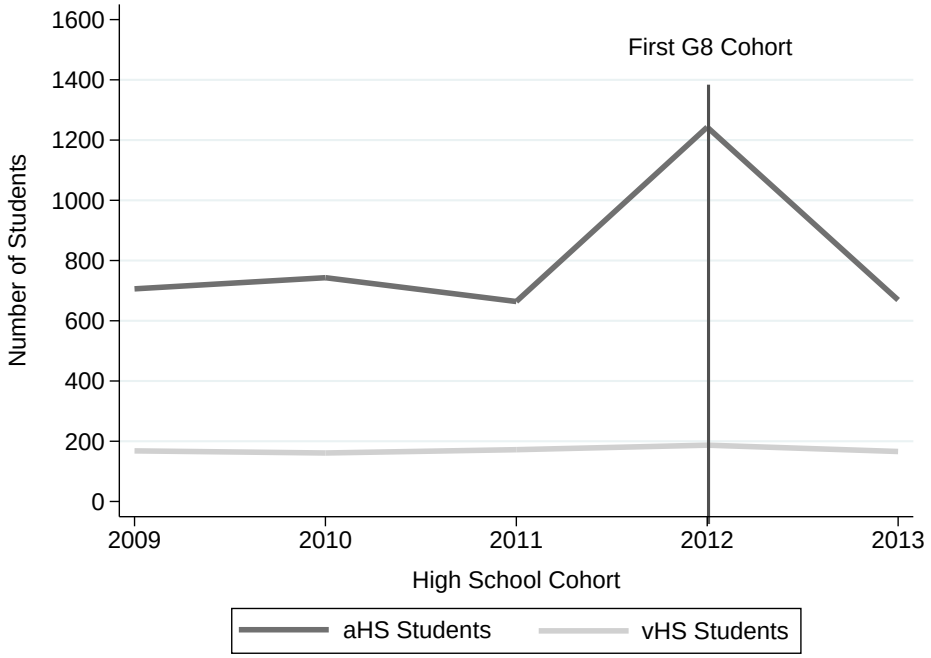
Table 2.1 presents summary statistics based on students of the 2009 to 2013 high school cohorts from Baden-Württemberg. The first column reports mean characteristics of the G8 students of the 2013 cohort; column two and three report mean characteristics of the 2012 G8 and G9 double cohort students; mean characteristics of the 2009 to 2011 academic high school graduates are shown in column four; characteristics of the 2009 to 2013 vocational high school graduates are reported in column five. Table 2.1 shows that students of the second G8 cohort obtained on average the highest grades while the G9 students from vocational high schools, i.e., the students of the control group, obtained on average the lowest grades. At the same time, the fraction of G8 students who took a

[11] We calculated the statewide ratio of vocational to academic high school graduates by using official data provided by the Statistical Office of the state of Baden-Württemberg. The ratio of vocational to academic high school graduates in our sample is calculated by using the numbers shown in Figure 2.1.

[12] We define the high school graduation date as the date of issue of the high school diploma, being issued latest in August of a given year.

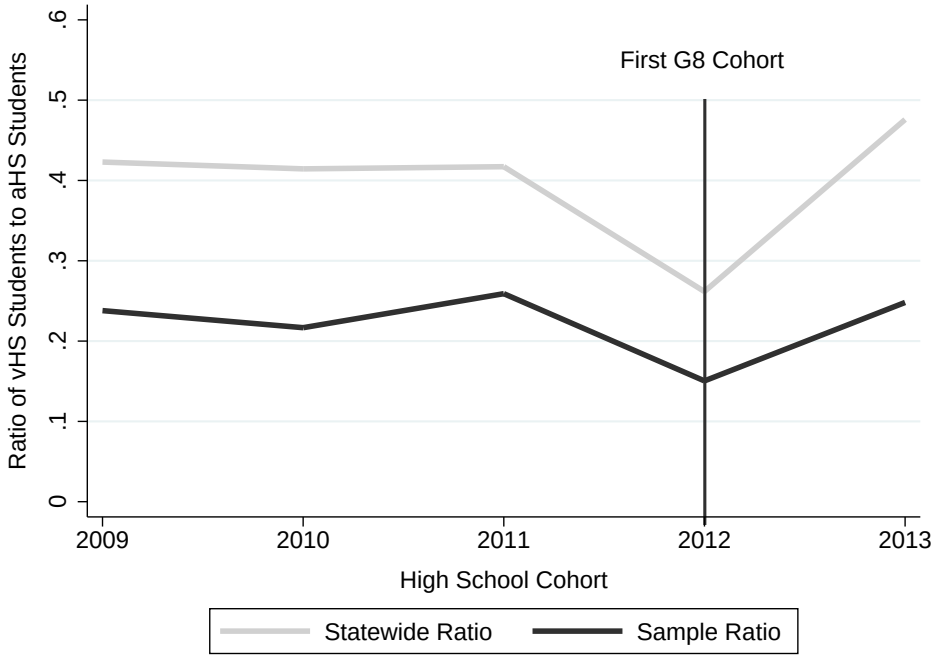
[13] The observation level of our data are exams; each student is represented according to the number of exams written.

Figure 2.1: Number of Students by High School Cohort



Note: aHS and *vHS* identify students who went to an academic or a vocational high school, respectively. Students who went to a G8 pilot school are excluded.

Figure 2.2: Ratio of vHS Students to aHS Students by High School Cohort



Note: Students who went to a G8 pilot school are excluded.

Table 2.1: Descriptive Statistics

HS Cohort	2013	2012		2009-11	2009-13
	G8	G8 DC	G9 DC	G9 - aHS	G9 - vHS
University grades	0.087	0.004	0.053	0.037	-0.226
	(1.009)	(0.929)	(0.980)	(0.939)	(0.940)
Break (in months)	11.683	10.776	9.894	10.393	10.464
	(7.734)	(7.580)	(7.157)	(7.258)	(7.613)
Break > one year	0.541	0.480	0.431	0.449	0.443
	(0.499)	(0.500)	(0.496)	(0.498)	(0.497)
School years	12.000	12.000	13.000	13.000	13.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
HS graduation age	18.461	18.480	19.449	19.459	19.432
	(0.262)	(0.282)	(0.275)	(0.285)	(0.282)
Male	0.457	0.476	0.502	0.460	0.424
	(0.499)	(0.500)	(0.500)	(0.499)	(0.495)
German	0.980	0.984	0.973	0.976	0.958
	(0.139)	(0.127)	(0.161)	(0.154)	(0.202)
#Students	505	550	564	1681	519

Note: Students who skipped a school year, were retained, went to a G8 pilot school, or enrolled later than two and a half years after their graduation are excluded. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Break* measures the time span between high school graduation and university enrollment. *aHS* and *vHS*, repsectively, identify students who went to an academic or a vocational high school. Standard errors are reported in parentheses.

one-year break after high school graduation is about five to ten percentage points higher than the fraction among the G9 students. By restricting the sample to students who were not retained, did not skip a grade, or went to a G8 pilot school, the G8 students are on average one year younger at the time of high school graduation and went to school for exactly one year less than the G9 students. The distribution of males as well as the share of German citizens is quite similar across the five groups. Overall, Table 2.1 shows that the treatment and comparison groups have fairly similar characteristics, and most importantly, there are hardly any changes over time in the characteristics of the two groups.

2.5 Empirical Strategy

To identify the effect of a reduction in years of schooling on academic achievement at the tertiary level, we exploit variation between school types over time within the state of Baden-Württemberg. Our approach is illustrated by Figure 2.3. Since 2012, students in Baden-Württemberg graduate from academic high schools after eight years, while the reform was not implemented at vocational high schools. Therefore, we can isolate the reform effect from cohort and school-type effects by estimating the following difference-in-differences model using the ordinary least squares (OLS) method:

$$Y_{ic}^{t} = \alpha \; HS\,Type_{i}^{t} + \beta \; Post_{ic} + \gamma \; Post_{ic} \times HS\,Type_{i}^{t} \; + \; \delta \; X_{i} \; + \; \eta_{ic}^{t}, \tag{2.1}$$

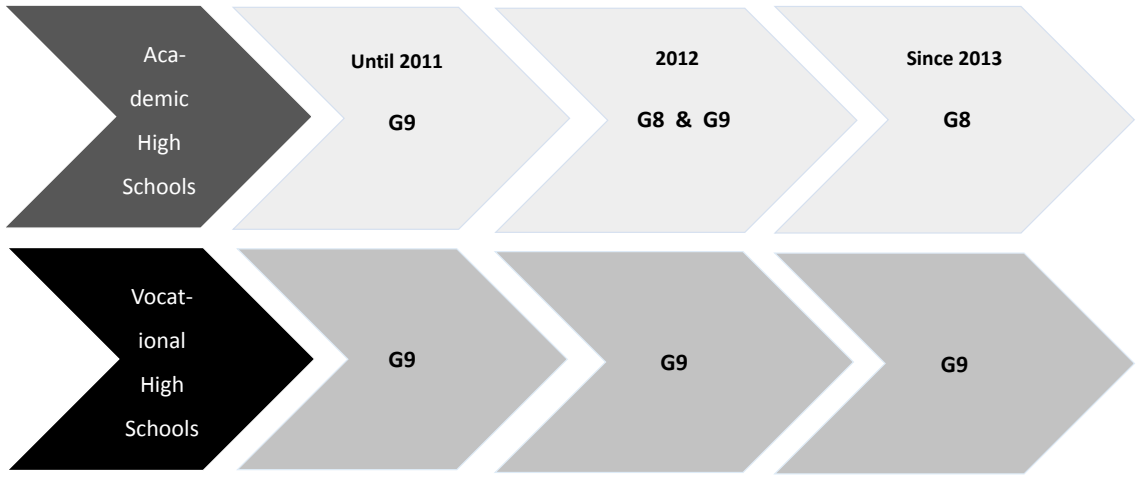
where Y_{ic}^{t} denotes the outcome of interest of student i of cohort c from school-type t. $HS\,Type_{i}^{t}$ is a dummy variable indicating whether a student graduated from an academic high school. The dummy variable $Post_{ic}$ indicates whether a student graduated from high school in 2012 or later. X_i is a vector of demographic and study-related covariates comprising dummies for sex, nationality, majors, semesters, and exams. Error terms η_{ic}^{t} are clustered at the individual level. Clustering the error terms at the individual level accounts for the presence of heteroscedasticity. Collapsing the time series information in a pre- and a postintervention period is a simple method to reduce the serial correlation problem (see Bertrand et al., 2004).

The coefficient of interest is γ, which is the difference-in-differences estimator of the impact of the one-year reduction in years of schooling on academic achievement in university. It measures the change in student achievement after the reform, relative to before the reform, among students who attended an academic high school relative to students who attended a vocational high school. The key identifying assumption is that there were no further changes in academic or vocational high schools in Baden-Württemberg simultaneously to the reform affecting the students' cognitive or non-cognitive skills, or the composition of the respective student bodies. Put differently, we assume that the underlying trends in the outcome variables would have been the same for both the treatment and the control group in the absence of the reform.

Figure 2.5 presents suggestive evidence for the validity of the common trend assumption (CTA). It plots the difference in the average university grade over time between academic and vocational high school students from Baden-Württemberg. The dashed lines represent the 95% confidence interval.[14] The double cohort graduated in Baden-Württemberg in 2012; from 2013 onwards, only G8 students graduated from academic high schools. The

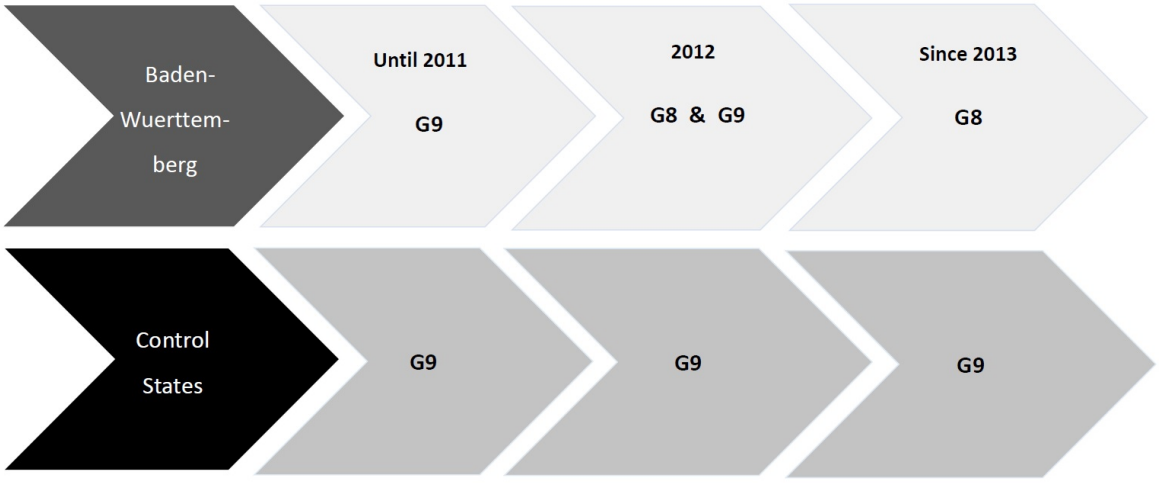
[14]The data are obtained by OLS regressions of Equation (2.1) without covariates using our baseline samples.

Figure 2.3: Identification Strategy 1: Between Schools, Across Time



Note: Only academic and vocational high school students from Baden-Württemberg considered.

Figure 2.4: Identification Strategy 2: Between States, Across Time

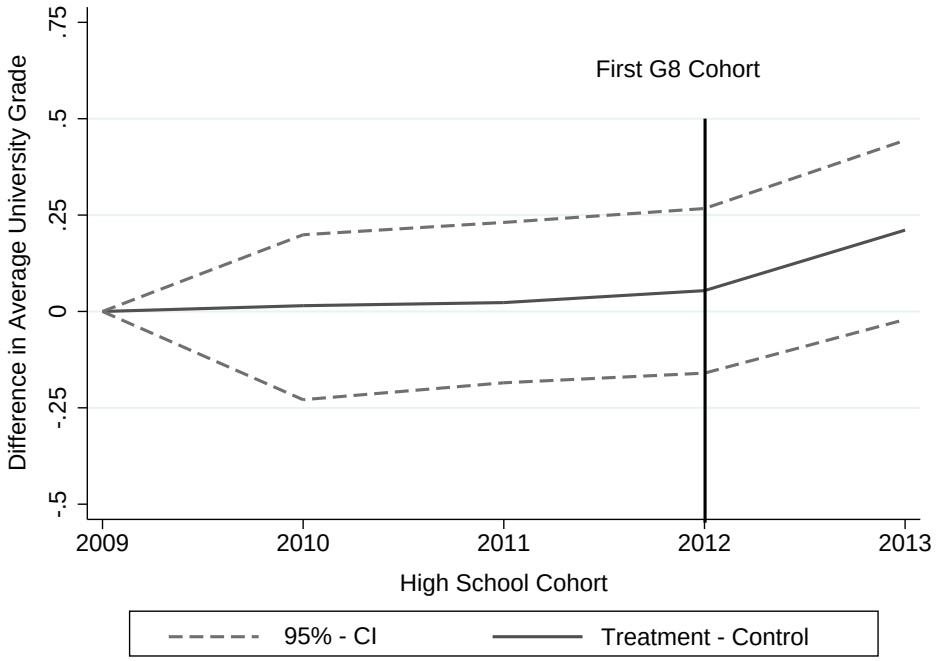


Note: Only academic high school students from Baden-Württemberg considered. Control group comprises academic high school students from Hesse, North Rhine-Westphalia, and Rhineland-Palatinate.

figure suggests that the CTA is fulfilled for Equation (2.1) as the difference in the average university grade of treatment and control students in the pre-treatment years is never significantly different from zero. However, the academic high school students of the 2012, and especially of the 2013 cohort outperformed the students of the control group. This is a first hint that the reform had rather a positive than a negative effect on university performance.

Our DiD analysis may still provide misleading estimates because of an omitted variables bias arising from unobserved and uncontrolled differences between the treatment and the

Figure 2.5: Difference in the Average University Grade between aHS Students (Treatment Group) and vHS Students (Control Group) by High School Cohort



Note: The solid line represents the difference in the average, standardized university grade between treatment and control students. The dashed-lines represent 95% confidence bands. The vertical line indicates the treatment year. Exams considered are those taken within the first two semesters and at the first attempt by more than 50 students. Students who skipped a school year, were retained, or went to a G8 pilot school are excluded.

control group. In particular, there are two main selection issues. First, if students opted less often for academic high schools to evade the reform, then Equation (2.1) will provide a biased estimate of the average causal effect of the reform on university performance. For instance, it could be that in particular weaker students opted more often for vocational high schools because they assumed that they would not be able to cope with the higher learning intensity. Figure 2.2, however, does not confirm this conjecture: the ratio of vocational to academic high school graduates in our sample follows the statewide ratio of vocational to academic high school graduates, and there is no increase in the total number of vocational high school graduates in the post-reform years.

Second, if the G8 reform had an effect on the university enrollment decision of academic high school graduates, then γ will differ from the average causal effect for the population of university-bound G8 students at large. For example, it could be that the more able and motivated students started their studies first. However, there is no evidence in our data for a different enrollment pattern among the students of the post-reform high school cohorts.

We further address the issue of an omitted variables bias in the robustness section, discussed after the main results, where we run a series of specification checks.

2.6 Results

In our discussion of the effects of the G8 reform on university performance, we first focus on the G8 double cohort students who graduated from academic high schools in Baden-Württemberg in 2012. Section 2.6.2 presents the effects of the reform on the students of the second G8 cohort who followed in 2013. In Section 2.6.3, we examine the sensitivity of our results to changes in the sample restrictions and the model specifications. Having demonstrated the robustness of our estimates, we investigate in Section 2.6.4 whether the effects of the reform vary in subgroups.

2.6.1 Effects on the First G8 Cohort

Table 2.2 presents results for the effect of the reform on students of the first G8 cohort, i.e., the double cohort students, with and without controls. The coefficient of the variable *G8-Reform Effect* represents γ, the parameter of interest in Equation (2.1). The model including controls is our baseline specification. The regressions are based on students of the 2009 to 2012 cohorts. The G9 double cohort students are excluded. The control group consists of those students who went to a vocational high school. As depicted in Figure 2.5, we do not find evidence for a violation of the CTA when estimating Equation (2.1) with a full set of interaction terms (see Table 2.A.1 in the appendix). The inclusion of the covariates does furthermore barely affect our estimates, supporting the assumption that the assignment to treatment and control group was random.

Table 2.2 reveals no statistically significant effect of the reform on the achievements of students of the first G8 cohort, although there is a slight positive tendency. Within the first two semesters, G8 students of the double cohort obtained grades that were on average slightly higher than the average grade obtained by the control students, and failed exams less often. The 95 percent confidence interval for the effect on the average grade, however, ranges from about minus 12 percent of a standard deviation to plus 24 percent of a standard deviation, while one fifth of a standard deviation is equivalent to a difference in the average grade of about one grading step. Thus, it is not possible to draw a clear-cut conclusion from this baseline estimate. The coefficient for the likelihood to obtain a top grade is basically zero, as well as the coefficient for the likelihood to drop out of university within the first two semesters. Considering the time span between high school graduation and university enrollment, our estimate suggest that the G8 double cohort students took, though not significantly, more often a one-year break after graduation than the control students. While only about 45 percent of the students of the G9 cohorts took a one-year break, this number increased to 48 percent among the G8 double cohort students.

Table 2.2: The Effect of the Reform on Students of the First G8 Cohort

Dependent variable	Average Grade		Failure Rate		Top Grade		Dropout Rate		One-Year Break	
G8-Reform Effect	0.050	0.063	-0.041	-0.035	0.011	-0.005	0.010	0.013	0.083	0.078
	(0.087)	(0.088)	(0.030)	(0.030)	(0.021)	(0.019)	(0.053)	(0.052)	(0.060)	(0.059)
Academic HS	0.183***	0.166***	-0.040***	-0.021	0.066***	0.038***	-0.078***	-0.056**	-0.007	-0.027
	(0.044)	(0.044)	(0.013)	(0.013)	(0.011)	(0.009)	(0.028)	(0.028)	(0.032)	(0.031)
Post Reform Cohort	-0.096	-0.117	0.068**	0.058**	-0.019	0.003	0.027	0.026	-0.049	-0.048
	(0.080)	(0.081)	(0.028)	(0.029)	(0.018)	(0.017)	(0.049)	(0.048)	(0.054)	(0.053)
Demographic and Study-Related Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No. of Students	2562	2562	2562	2562	2562	2562	2562	2562	2562	2562
Exam Observations	21870	21870	21870	21870	21870	21870				

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

Note: OLS regressions are based on 2009-2012 high school cohorts from Baden-Württemberg. The control group consists of vocational high school students from Baden-Württemberg. The G9 double cohort students as well as students who skipped a school year, were retained, went to a G8 pilot school, or enrolled later than two and a half years after their graduation are excluded. Exams considered are those taken within the first two semesters at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Academic HS* and *Post Reform Cohort* are equal to one if a student attended an academic high school, or belonged to the post-reform cohort, respectively. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

2.6.2 Effects on the Second G8 Cohort

Table 2.3 presents estimates of the effect of the reform for students of the second G8 cohort, with and without controls. The model including controls represents our baseline specification. Regressions are based on the 2009 to 2013 cohorts, whereas students of the 2012 cohort are excluded because of the double cohort. The control group consists of students who graduated from a vocational high school. Again, there is no evidence for a violation of the CTA when estimating Equation (2.1) with a full set of interaction terms (see Table 2.A.2 in the appendix). The inclusion of the control variables does also not significantly affect our estimates, supporting the assumption that the assignment to treatment and control group was random.

Table 2.3 shows that students of the second G8 cohort obtained grades that were on average about one fifth of a standard deviation higher than the average grade obtained by the control students. This effect is statistically significant at the five percent level, and equivalent to a difference in the average grade of about one grading step. Considering the 95 percent confidence interval, the effect ranges from basically a zero effect to a difference in the average grade of almost two grading steps. Thus, the tendency is clearly positive. The G8 students of the second cohort were also by about eight percentage points less likely to fail an exam. This effect is also statistically significant at the five percent level. The coefficient for the likelihood to drop out of university within the first two semesters is slightly positive, but not statistically significant. The pattern of a slightly delayed enrollment also shows up for the students of the second G8 cohort: Our estimate suggests that G8 students were about nine percentage points more likely to take a one-year break after graduation, although the coefficient is not statistically significant. Overall, the results suggest that the reform had on average a positive effect on the academic achievement of students of the second G8 cohort.

2.6.3 Robustness Checks

In the following, we show that the results presented above are robust to a variety of alternative specification choices and validity checks. In particular, we perform placebo tests, check whether our results differ if we consider a different control group, or a different treatment group, or when we vary the break restriction, the exam restriction, or the cut-off date for school enrollment.

If the effects are causal effects of the one-year reduction in years of schooling, effects should be present for students who graduated from academic high schools in Baden-Württemberg in 2012 and later with no corresponding significant effects for academic high school students of earlier cohorts. Columns two and three of Table 2.4 show results of

Table 2.3: The Effect of the Reform on Students of the Second G8 Cohort

Dependent variable	Average Grade		Failure Rate		Top Grade		Dropout Rate		One-Year Break	
G8-Reform Effect	0.232**	0.228**	-0.088***	-0.082***	0.016	0.022	-0.022	0.052	0.102	0.085
	(0.091)	(0.091)	(0.032)	(0.032)	(0.018)	(0.018)	(0.057)	(0.056)	(0.063)	(0.062)
Academic HS	0.175***	0.155***	-0.033**	-0.015	0.064***	0.035***	-0.070**	-0.046	-0.000	-0.015
	(0.045)	(0.035)	(0.013)	(0.013)	(0.010)	(0.007)	(0.029)	(0.029)	(0.033)	(0.031)
Post Reform Cohort	-0.200**	-0.215**	0.115***	0.112***	-0.016	-0.019	0.048	0.052	-0.018	-0.013
	(0.083)	(0.065)	(0.030)	(0.025)	(0.016)	(0.012)	(0.053)	(0.052)	(0.058)	(0.056)
Demographic and Study-Related Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No. of Students	2326	2326	2326	2326	2326	2326	2326	2326	2326	2326
Exam Observations	19963	19963	19963	19963	19963	19963				

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Note: OLS regressions are based on 2009-2013 high school cohorts from Baden-Württemberg, while the 2012 cohorts are excluded. The control group consists of vocational high school students from Baden-Württemberg. Students who skipped a school year, were retained, went to a G8 pilot school, or enrolled later than one and a half years after their graduation are excluded. Exams considered are those taken within the first two semesters at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Academic HS* and *Post Reform Cohort* are equal to one if a student attended an academic high school, or belonged to the post-reform cohort, respectively. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

placebo tests in the timing of the reform. In the second column, the placebo treatment group consists of the students of the 2011 academic high school cohort of Baden-Württemberg. In the third column, the placebo treatment group are the students of the 2010 academic high school cohort of Baden-Württemberg. The control group consists of the vocational high school students of the respective cohorts. The estimates show that there is virtually no difference in the achievement measures between treatment and control students in the pretreatment years. At most, the results suggest that the academic high school students of the 2011 cohort took slightly less often a one-year break. However, the effect size is relatively small considering that about 55 percent of the students of a high school cohort start their studies immediately after their graduation. The coefficient is furthermore not statistically different from zero at the ten percent level. Thus, the placebo tests support a causal interpretation of our estimates.

Another concern may be that students who went to a vocational high school are not an appropriate control group, e.g., because students may have selected into vocational high schools to evade the G8 reform. The timing of the implementation of the reform, however, also allows us to exploit variation between states over time. Thus, we can check this concern by using academic high school students from other states as a control group. Among the freshmen students, every academic year there are about seven percent who obtained their high school diploma in North Rhine-Westphalia, or Rhineland-Palatinate. In both states, only G9 students graduated from academic high schools in 2012. This allows us to use these students as a control group for estimating the effect of the reform on the G8 double cohort students of Baden-Württemberg. The results are reported in column four of Table 2.4. Column one reports the results of our baseline specification using the vocational high school students from Baden-Württemberg as the control group. The estimates for the effect of the reform on academic achievement are very similar or even identical, suggesting that the results of our baseline specification are not driven by the choice of the control group. However, there is a significant difference concerning the effect of the reform on the students' enrollment decision measured by the outcome variable "One-year break". Using the academic high school students from North Rhine-Westphalia and Rhineland-Palatinate as the control group, the estimate implies that the share of G8 students who took at least a one-year break after high school graduation is about one quarter higher than before. In fact, this increase is driven by the control group. While about 55 percent of all students of the 2009 to 2011 high school cohorts in our sample enrolled in the year in which they graduated from high school, this number jumps to 69 percent for the students of the 2011 cohort, and to 75 percent for the students of the 2012 cohort of North Rhine-Westphalia and Rhineland-Palatinate. Therefore, the academic high school students from North Rhine-Westphalia and Rhineland-Palatinate do not seem

Table 2.4: Robustness Checks: First Cohort

	Baseline	Placebo -1 Year	Placebo -2 Years	Control Group	Treatm. Group	No Break Restrict.	1-Year Break	> 25 Students	> 50 Students	Birthday Cut-off
Dependent variable										
Average Grade	0.063	0.027	-0.022	0.021	0.036	0.058	0.098	0.054	0.056	0.035
	(0.088)	(0.088)	(0.117)	(0.106)	(0.163)	(0.085)	(0.091)	(0.091)	(0.094)	(0.093)
Failure Rate	-0.035	0.036	-0.014	-0.014	-0.035	-0.033	-0.054*	-0.033	-0.045	-0.039
	(0.030)	(0.027)	(0.036)	(0.032)	(0.041)	(0.030)	(0.032)	(0.032)	(0.033)	(0.032)
Top Grade	-0.005	0.026	-0.021	0.034	-0.026	-0.008	-0.003	-0.011	-0.013	-0.013
	(0.019)	(0.018)	(0.024)	(0.027)	(0.047)	(0.018)	(0.020)	(0.019)	(0.019)	(0.020)
Dropout Rate	0.013	-0.043	0.007	0.009	0.007	0.013	-0.014	0.033	0.050	0.016
	(0.052)	(0.058)	(0.066)	(0.053)	(0.075)	(0.053)	(0.055)	(0.053)	(0.055)	(0.056)
One-Year Break	0.078	-0.077	0.023	0.246***	0.200*			0.094	0.093	0.110*
	(0.059)	(0.063)	(0.073)	(0.071)	(0.111)			(0.060)	(0.061)	(0.063)
No. of Students	2562	1901	1283	2379	370	2713	2378	2417	2256	2135
Exam Observations	21870	16361	11156	20676	3337	22823	20444	18702	16644	18394

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

Note: Students who skipped a school year, were retained, or went to a G8 pilot school are excluded. Exams considered are those taken within the first two semesters and at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

to be an appropriate control group for evaluating the effect of the reform on the students' enrollment decision. For estimating the effect on academic achievement, however, our additional robustness checks show the appropriateness of this control group.

Similar to the concern regarding the control group, one may worry that the results differ for G8 students from other states. We check this potential issue by considering academic high school students from Bavaria as another treatment group. Each academic year, about eight percent of the freshman students obtained their high school diploma in Bavaria which is a neighboring state of Baden-Württemberg. Bavaria introduced the reform in the school year 2004/5. The G8 students of the double cohort graduated from academic high schools in 2011; the students of the second G8 cohort followed in 2012. Using the academic high school students from North Rhine-Westphalia and Rhineland-Palatinate as the control group, we can estimate Equation (2.1) to evaluate the effect of the reform on the Bavarian G8 students. Column five of Table 2.4 presents the results for the first G8 cohort, based on the students of the 2009 to 2011 high school cohorts. The G9 double cohort students are excluded in these regressions. Column two of Table 2.5 presents the results for the second G8 cohort, based on the 2009 to 2012 cohorts. In these regressions, the 2011 high school cohort is excluded. The results strongly support the validity of our main specification; the effects on the Bavarian G8 students are quantitatively and qualitatively very similar to the estimated effects on the Baden-Württemberg G8 students, for both the first and the second cohort. The much higher coefficients for the "One-year break" variable is again explained by the unusually high share of students of the control group who enrolled immediately after graduation in 2011 and 2012.

The maximum break a G8 student of the first cohort in our sample could take is two and a half years. To increase the comparability of our treatment and control group, we therefore restrict the sample for estimating the effect of the reform on the first G8 cohort to students who enrolled at the University of Konstanz at most 30 months after graduating from high school. For the same reason, we restrict the sample for estimating the effect on the second G8 cohort to students who enrolled at most 18 months after graduating from high school. Especially the latter restriction may harm the external validity of our results for the second cohort if achievements differ significantly between students who enroll latest one and a half years after graduation, and students who enroll more than one and a half years after graduation. Columns six and seven of Table 2.4 show the average treatment effect on the students of the first G8 cohort when altering the break restriction. The estimates confirm that the result from our baseline specification is very robust with respect to changes in the break duration. In particular, the results for the first cohort should be applicable to the second one as about 96 percent of the university-bound students of a cohort enroll at most two and a half years after graduation from high school. Thus, we

Table 2.5: Robustness Checks: Second Cohort

	Baseline	Treatm. Group	> 25 Students	> 50 Students
Dependent variable				
Average Grade	0.228**	0.252	0.209**	0.202**
	(0.091)	(0.173)	(0.094)	(0.096)
Failure Rate	-0.082***	-0.063	-0.075**	-0.073**
	(0.032)	(0.050)	(0.031)	(0.032)
Top Grade	0.022	0.035	0.012	0.014
	(0.018)	(0.045)	(0.018)	(0.018)
Dropout Rate	0.052	-0.039	-0.030	-0.022
	(0.056)	(0.086)	(0.057)	(0.060)
One-Year Break	0.085	-0.249***	0.079	0.082
	(0.062)	(0.075)	(0.063)	(0.066)
No. of Students	2326	319	2203	2058
No. of Exam Observations	19963	2936	17155	15304

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

Note: Students who skipped a school year, were retained, or went to a G8 pilot school are excluded. Exams considered are those taken within the first two semesters and at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

have no reason to believe that the positive effect we find for the G8 students of the second cohort is driven by the fact that the better G8 students enroll earlier.

One of the main advantages of our data is that we observe academic achievement of treatment and control students obtained in the same exams. In our baseline specification, we make no restriction concerning the size of an exam, measured as the number of students taking it. However, a conjecture may be that the grading differs between exams taken by 10 students, and exams taken by 50 or more students, for example, because the assessment of a single exam may become more objective the more exams are assessed jointly. Restricting the exam size, for instance, to at least 50 students makes it additionally more likely that we observe exams taken by both treatment and control students, increasing the comparability of the grades obtained. Columns eight and nine of Table 2.4 as well as columns three and four of Table 2.5 present the average treatment effect on the students of the first and the second G8 cohort, respectively, when altering the exam size. The estimates confirm that our baseline results are very robust to the use of these alternative specifications.

A final issue is the assignment of the students of the double cohort to a G8 or a G9 cohort which relies on the cut-off date for school enrollment (30th of June), and consequently on the dates of birth of the students. For our estimates it is crucial that the students who were born around the cut-off date were not enrolled earlier than usually to evade the reform.[15] As described in Section 2.4, we conducted a survey among all currently enrolled undergraduate students to validate our assignment, and the results provide no evidence for a significant number of misassignments. Additionally, we can check the robustness of our baseline specification by restricting the sample to students who were not born close to the cut-off date. As column ten of Table 2.4 shows, our estimates barely change when excluding the students who were born in July and August. Thus, we have no reason to believe that we systematically misassigned students of the double cohort to a G8 or G9 cohort.

2.6.4 Subgroup Analysis

Our analysis thus far has focused on the average effect of the German high school reform on the academic achievement of university students, finding significant differences between treatment and control students. Additionally, there could be important heterogeneity in the treatment effect across subgroups. For example, Büttner and Thomsen (2013) and Andrietti (2015) find differing effects with respect to gender, ability, and subjects. Tables 2.6 and 2.7 present estimates of the treatment effect for the students of the first and the second G8 cohort, respectively, when dividing our sample into these subgroups.

The first two columns of Tables 2.6 and 2.7 provide evidence for heterogeneous effects with respect to gender. We find slightly positive, but insignificant effects on the academic achievement of the male students of both G8 cohorts, as well as on the academic achievement of the female students of the first G8 cohort. For the female G8 students of the second cohort, we find significant positive effects: Their average grade is almost one third of a standard deviation higher than the average grade of the female G9 students, which corresponds to an improvement of about one grading step. The female G8 students of the second cohort were also ten percentage points less likely to fail an exam. Both effects are also significantly different from the effects on the male students. Concerning the time span between high school graduation and university enrollment, our estimates show that the female G8 students of the second cohort took significantly more often a one-year break after graduation than the students of the control group. For the male students we find no significant difference.

The reform may also have affected higher and lower ability students differently. By

[15] In Section 2.5, we discuss that exceptional school enrollments due to the reform were very unlikely, and can also not be observed in official statistics.

Table 2.6: Heterogeneous Effects: First Cohort

	w.r.t. the Gender		w.r.t. the HSGPA		w.r.t. the Subject	
	Male	Female	Top 50%	Bottom 50%	Math.	Non-Math.
Dependent variable						
Average Grade	0.022	0.099	0.128	0.069	0.038	0.093
	(0.143)	(0.111)	(0.112)	(0.107)	(0.112)	(0.122)
Failure Rate	-0.031	-0.029	-0.033	-0.064	-0.040	-0.014
	(0.052)	(0.037)	(0.031)	(0.051)	(0.042)	(0.032)
Top Grade	-0.000	0.001	-0.006	-0.011	-0.013	0.018
	(0.027)	(0.028)	(0.030)	(0.015)	(0.022)	(0.033)
Dropout Rate	0.087	-0.040	0.056	-0.048	0.004	0.048
	(0.081)	(0.064)	(0.066)	(0.082)	(0.065)	(0.084)
One-Year Break	0.115	0.080	0.062	0.088	0.109	-0.011
	(0.089)	(0.078)	(0.079)	(0.088)	(0.070)	(0.107)
No. of Students	1187	1375	1303	1259	1556	1000
Exam Observations	9616	12254	11644	10226	11822	9945

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

Note: OLS regressions are based on 2009-2012 high school cohorts from Baden-Württemberg. The control group consists of vocational high school students from Baden-Württemberg. The G9 double cohort students as well as students who skipped a school year, were retained, went to a G8 pilot school, or enrolled later than two and a half years after their graduation are excluded. Exams considered are those taken within the first two semesters and at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

subtracting the average statewide high school grade point average (HSGPA) from a student's HSGPA (performed for each cohort of academic and vocational high school students separately), we get a measure for a student's relative ability within his cohort. Columns three and four of Tables 2.6 and 2.7 show the results when running regressions for the students belonging to the lower 50 percent of their high school cohort, and the students belonging to the upper 50 percent of their high school cohort with respect to their HSGPA. In terms of the average grade, our results indicate that the higher ability students of the first G8 cohort improved more than the lower ability students, though both effects are individually not statistically significantly different from zero. Our estimates for the second G8 cohort indicate that the average grade of both the lower and the higher ability students improved by about one fifth of a standard deviation. Considering the bottom tail of the grade distribution, especially the lower ability students of the second G8 cohort failed an

Table 2.7: Heterogeneous Effects: Second Cohort

	w.r.t. the Gender		w.r.t. the HSGPA		w.r.t. the Subject	
	Male	Female	Top 50%	Bottom 50%	Math.	Non-Math.
Dependent variable						
Average Grade	0.060	0.312***	0.216*	0.230**	0.210*	0.224
	(0.152)	(0.114)	(0.115)	(0.106)	(0.114)	(0.148)
Failure Rate	-0.039	-0.103***	-0.062*	-0.098**	-0.081*	-0.069*
	(0.057)	(0.038)	(0.035)	(0.047)	(0.042)	(0.041)
Top Grade	0.026	0.012	0.030	0.009	0.017	0.037
	(0.030)	(0.024)	(0.028)	(0.012)	(0.019)	(0.040)
Dropout Rate	0.019	-0.042	0.064	-0.099	-0.047	0.021
	(0.089)	(0.072)	(0.068)	(0.088)	(0.071)	(0.090)
One-Year Break	0.031	0.155*	0.055	0.099	0.023	0.227**
	(0.091)	(0.083)	(0.085)	(0.091)	(0.076)	(0.105)
No. of Students	1055	1271	1198	1128	1437	877
Exam Observations	8567	11396	10890	9073	11035	8829

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

Note: OLS regressions are based on 2009-2013 high school cohorts from Baden-Württemberg, while the 2012 cohorts are excluded. The control group consists of vocational high school students from Baden-Württemberg. Students who skipped a school year, were retained, went to a G8 pilot school, or enrolled later than one and a half years after their graduation are excluded. Exams considered are those taken within the first two semesters and at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

exam considerably less often than the students of the control group. The coefficient for the lower ability students of the first G8 cohort is also negative, but insignificant. The higher ability academic high school students failed an exam already before the reform in only about seven percent of the cases. Correspondingly low and statistically insignificant are the coefficients. Overall, our estimates suggest that the reform has improved the academic achievement of both low- and high-achieving students with slightly stronger effects for the low-achieving students.

The evidence concerning heterogeneous effects with respect to subjects is mixed so far. While Büttner and Thomsen (2013) find negative effects on high school grades in mathematics, and no effect on grades in German literature, Andrietti (2015) reports positive effects on the reading, mathematics, and science literacy skills of high school students. We address this topic by considering classes with a mathematical content, and

those with a non-mathematical content separately. We define mathematics, computer sciences, information engineering, biology, chemistry, physics, life science, nano science, economics, and mathematical finance as classes with a math-heavy content. Social sciences, political sciences, sports, psychology, languages, language sciences, literature sciences, history, and arts are defined as classes with a non-mathematical content. The results are shown in columns five and six of Tables 2.6 and 2.7. For the first cohort, the effects are small and insignificant for both types of classes considered. The effects are larger for the second cohort, and also statistically significant for some outcomes. Splitting the samples of the second cohort further with respect to gender, we find that the rather positive effects in mathematical classes stem from the female students, while the coefficients for the male students are close to zero. In the non-mathematical classes, the achievement of G8 students of both genders has seemingly improved, though the coefficients are not statistically significant because of the then too small sample sizes. The increase in the likelihood to take a one-year break is solely driven by the female students.

2.7 Discussion and Conclusion

This article investigates a recent high school reform implemented in Germany that reduced the duration of high school from nine to eight years but left the curriculum as well as the total instruction time unchanged. Thus, the reform compressed the established program by one year. Using student-level data of the University of Konstanz, we estimate the impact of this reduction of the years of schooling on university students' enrollment decisions and academic achievement, i.e., longer-term effects. Since the states carried out the reform in different years and implemented it only in academic high schools, we can disentangle the reform effect from cohort, state, and school-type effects by a DiD strategy.

The estimates of our baseline specification show no effect for students of the first G8 cohort. For students of the second G8 cohort, we find significant positive effects. Thus, affected students obtained grades that were on average one fifth of a standard deviation higher (about one grading step) than the grades of the control group. Affected students were also less likely to fail an exam. Several robustness checks support these findings. Additionally, we find significant heterogeneity in the treatment effect in terms of gender, ability, and class content. Thus, the positive effects we find for students of the second G8 cohort stem particularly from the female students. The effects are also slightly more positive for weaker students. At the same time, we find no evidence for a negative effect on learning and human capital accumulation for any subgroup. Thus, our results suggest that there may be scope for a reduction of the years of schooling if curricula and total instruction time are not altered.

One explanation for the positive effects we find may be that the reform increased the requirements for high school students, thereby preparing them better for university studies. G8 students were forced to develop better learning strategies in school, and got used to a higher learning intensity and a higher stress level. In particular female students seem to succeed in coping with these higher requirements.[16]

The heterogeneity in the treatment effect between genders may be explained by the fact that females show on average less problematic behavior in high school, are more self-disciplined, and thus, better able to cope with stress. For example, Fischer et al. (2013) find that females show more compensatory effort and self-control, and take more pride in their own productivity which helps them to outperform their male counterparts in secondary school. Taylor et al. (2000) further show that when it comes to stress, women become more likely to express affiliative social behavior, either to befriend the enemy - if there is an enemy and is causing the stress - or to seek social support from their family members or friends. Although fight-or-flight may characterize the primary physiological responses to stress for both males and females, Taylor et al. (2000) propose that, behaviorally, females' responses are more marked by a pattern of "tend-and-befriend".

One caveat, however, has to be kept in mind: The study by Huebener and Marcus (2017) shows that some students are left behind by the reform and have to repeat a grade. Thus, the most poorly performing students may not be able to cope with the increased requirements. Our data support the finding by Huebener and Marcus (2017), as the fraction of repeaters in the second G8 cohort is higher than it was before the reform.[17] However, our findings barely change when we include students who repeated a grade once in our analysis. In particular, we find no evidence that repeaters of the second G8 cohort perform worse than those of earlier G9 cohorts.

In sum, our estimates provide robust evidence that the reform has no detrimental effect on academic achievement of university students. Instead, the reform reduces the opportunity costs of high school education and facilitates an earlier labor market entry as students are on average almost one year younger when they leave school.

[16]This pattern is also found by Andrietti (2015) using PISA data.

[17]Our sample of students belonging to the first G8 cohort only consists of G8 students with a regular school career. Those G8 students of the first cohort who had to repeat a grade once, ended up in the second G8 cohort. Those who originally started in the first G8 cohort and repeated a grade twice would be in the third G8 cohort in our regressions.

References

Anderson, D. M. and Walker, M. (2015). Does shortening the school week impact student performance? Evidence from the four-day school week. *Education Finance and Policy*, 10(3):314–349.

Andrietti, V. (2015). The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment. Universidad Carlos III de Madrid Working Paper Economic Series 15-06.

Battistin, E. and Meroni, E. C. (2016). Should we increase instruction time in low achieving schools? Evidence from Southern Italy. *Economics of Education Review*, 55:39 – 56.

Bellei, C. (2009). Does lengthening the school day increase students academic achievement? Results from a natural experiment in Chile. *Economics of Education Review*, 28(5):629–640.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Büttner, B. and Thomsen, S. L. (2013). Are we spending too many years in school? Causal evidence of the impact of shortening secondary school duration. *German Economic Review*, 16(1):65–86.

Card, D. (1999). The Causal Effect of Education on Earnings. In Ashenfelter, O. C. and Card, D., editors, *Handbook of Labor Economics*, volume 3, pages 1801 – 1863. Amsterdam: Elsevier.

Carneiro, P., Crawford, C., and Goodman, A. (2007). The impact of early cognitive and non-cognitive skills on later outcomes. CEE Discussion Papers, Centre for the Economics of Education, LSE.

Dahmann, S. and Anger, S. (2014). The impact of education on personality: Evidence from a German high school reform. SOEPPapers on Multidisciplinary Panel Data Research 658, DIW Berlin.

Dahmann, S. C. (2017). How does education improve cognitive skills? Instructional time versus timing of instruction. *Labour Economics*.

Eren, O. and Millimet, D. L. (2007). Time to learn? The organizational structure of schools and student achievement. *Empirical Economics*, 32(2-3):301–332.

Fischer, F., Schult, J., and Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*, 28(2):529–543.

Hansen, B. (2011). School year length and student performance: Quasi-experimental evidence. Technical report.

Hanushek, E., Rivkin, S., and Taylor, L. L. (1996). Aggregation and the estimated effects of school resources. *The Review of Economics and Statistics*, 78(4):611–27.

Hanushek, E. A. and Wößmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3):607–68.

Homuth, C. (2012). Der Einfluss des achtjährigen Gymnasiums auf den Kompetenzerwerb. Technical report.

Huebener, M. and Marcus, J. (2017). Compressing instruction time into fewer years of schooling and the impact on student performance. *Economics of Education Review*, 58:1 – 14.

Klemm, K. (2008). Bildungszeit: Vom Umgang mit einem knappen Gut. In *Schulzeiten, Lernzeiten, Lebenszeiten*, pages 21–30. Zeiher, H./ Schroeder, S., Weinheim.

Krashinsky, H. (2014). How would one extra year of high school affect academic performance in university? Evidence from an educational policy change. *Canadian Journal of Economics*, 47(1):70–97.

Kühn, S. M., van Ackeren, I., Bellenberg, G., Reintjes, C., and im Brahm, J.-P. D. G. (2013). Wie viele Schuljahre bis zum Abitur? *Zeitschrift für Erziehungswissenschaft*, 16(1):115–136.

Lavy, V. (2012). Expanding school resources and increasing time on task: Effects of a policy experiment in Israel on student academic achievement and behaviour. NBER Working Paper 18369.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, 125(588):F397–F424.

Lee, J.-W. and Barro, R. J. (2001). Schooling quality in a cross–section of countries. *Economica*, 68(272):465–488.

Lehn, B. v. (2010). *Generation G8. Wie die Turbo-Schule Schüler und Familien ruiniert.* Weinheim: Beltz.

Lochner, L. (2011). Nonproduction benefits of education: Crime, health, and good citizenship. *Handbook of the Economics of Education*, 4:183.

Mandel, P. and Süssmuth, B. (2011). Total instructional time exposure and student achievement: An extreme bounds analysis based on German state-level variation. CESifo Working Paper Series 3580.

Marcotte, D. E. (2007). Schooling and test scores: A mother-natural experiment. *Economics of Education Review*, 26(5):629–640.

Marcotte, D. E. and Hemelt, S. W. (2008). Unscheduled school closings and student performance. *Education Finance and Policy*, 3(3):316–338.

Meyer, T. and Thomsen, S. L. (2013). Are 12 years of schooling sufficient preparation for tertiary education? Evidence from the reform of secondary school duration in Germany. NIW Discussion Paper 8.

Meyer, T. and Thomsen, S. L. (2016). How important is secondary school duration for postsecondary education decisions? Evidence from a natural experiment. *Journal of Human Capital*, 10(1):67–108.

Meyer, T., Thomsen, S. L., and Schneider, H. (2015). New Evidence on the Effects of the Shortened School Duration in the German States: An Evaluation of Post-Secondary Education Decisions. IZA Discussion Papers 9507, Institute for the Study of Labor (IZA).

Morin, L.-P. (2013). Estimating the benefit of high school for universitybound students: Evidence of subjectspecific human capital accumulation. *Canadian Journal of Economics*, 46(2):441–468.

OECD (2014). *Education at a Glance 2014: OECD Indicators*. OECD Publishing.

Patall, E. A., Cooper, H., and Allen, A. B. (2010). Extending the school day or school year a systematic review of research (1985–2009). *Review of Educational Research*, 80(3):401–436.

Pischke, J.-S. (2007). The impact of length of the school year on student performance and earnings: Evidence from the German short school years. *The Economic Journal*, 117(523):1216–1242.

Quis, J. S. and Reif, S. (2017). Health effects of instruction intensity. Evidence from a natural experiment in german high-schools. BERG Working Paper Series 123, Bamberg University.

Statistisches Bundesamt (2014). Bildung und Kultur: Allgemeinbildende Schulen. Fachserie 11 Reihe 1, Statistisches Bundesamt, Wiesbaden.

Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A., and Updegraff, J. A. (2000). Biobehavioral Responses to Stress in Females: Tend-and-Befriend, Not Fight-or-Flight. *Psychological Review*, 107(3):411–429.

Thiel, H., Thomsen, S. L., and Büttner, B. (2014). Variation of learning intensity in late adolescence and the effect on personality traits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(4):861–892.

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2):117–170.

2.8 Appendix

Table 2.A.1: The Effect of the Reform on Students of the First G8 Cohort, Extended

Dependent variable	Average Grade	Failure Rate	Top Grade	Dropout rate	One-Year Break
G8-Reform Effect	0.062	-0.027	-0.008	0.006	0.079
	(0.114)	(0.038)	(0.023)	(0.066)	(0.073)
Academic HS	0.166**	-0.030	0.041***	-0.049	-0.025
	(0.084)	(0.026)	(0.016)	(0.039)	(0.053)
Cohorts					
2010	0.045	-0.006	0.035*	0.001	-0.162**
	(0.107)	(0.032)	(0.021)	(0.062)	(0.068)
2011	0.028	-0.041	0.007	0.038	-0.237***
	(0.100)	(0.031)	(0.018)	(0.064)	(0.068)
2012	-0.090	0.042	0.018	0.039	-0.188***
	(0.107)	(0.036)	(0.020)	(0.061)	(0.067)
Interaction Terms					
Academic HS x 2010	-0.015	-0.008	-0.021	0.010	0.038
	(0.113)	(0.034)	(0.023)	(0.066)	(0.074)
Academic HS x 2011	0.016	0.033	0.013	-0.031	-0.056
	(0.107)	(0.034)	(0.021)	(0.068)	(0.073)
Demographic and Study-Related Covariates	Yes	Yes	Yes	Yes	Yes
No. of Students	2562	2562	2562	2562	2562
Exam Observations	21870	21870	21870		

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

Note: OLS regressions are based on 2009-2012 high school cohorts from Baden-Württemberg. The control group consists of vocational high school students from Baden-Württemberg. The G9 double cohort students as well as students who skipped a school year, were retained, went to a G8 pilot school, or enrolled later than two and a half years after their graduation are excluded. Exams considered are those taken within the first two semesters at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Academic HS* is equal to one if a student attended an academic high school. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

Table 2.A.2: The Effect of the Reform on Students of the Second G8 Cohort, Extended

Dependent variable	Average Grade	Failure Rate	Top Grade	Dropout Rate	One-Year Break
G8-Reform Effect	0.247**	-0.080**	0.019	-0.020	0.095
	(0.117)	(0.039)	(0.023)	(0.070)	(0.077)
Academic HS	0.136	-0.017	0.038**	-0.047	-0.025
	(0.086)	(0.026)	(0.017)	(0.051)	(0.055)
Cohorts					
2010	-0.012	0.011	0.035	-0.014	-0.197***
	(0.111)	(0.033)	(0.023)	(0.064)	(0.069)
2011	-0.007	-0.038	0.006	0.031	-0.238***
	(0.103)	(0.030)	(0.019)	(0.066)	(0.069)
2013	-0.221**	0.103***	-0.003	0.058	-0.164**
	(0.109)	(0.037)	(0.020)	(0.065)	(0.070)
Interaction Terms					
Academic HS x 2010	0.023	-0.023	-0.023	0.030	0.080
	(0.118)	(0.035)	(0.025)	(0.068)	(0.075)
Academic HS x 2011	0.034	0.030	0.013	-0.025	-0.063
	(0.110)	(0.033)	(0.022)	(0.070)	(0.074)
Demographic and Study-Related Covariates	Yes	Yes	Yes	Yes	Yes
No. of Students	2326	2326	2326	2326	2326
Exam Observations	19963	19963	19963		

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

Note: OLS regressions are based on 2009-2013 high school cohorts from Baden-Württemberg, while the 2012 cohorts are excluded. The control group consists of vocational high school students from Baden-Württemberg. Students who skipped a school year, were retained, went to a G8 pilot school, or enrolled later than one and a half years after their graduation are excluded. Exams considered are those taken within the first two semesters at the first attempt. Grades are standardized by exam level to have a mean of zero and a standard deviation of one. *Failure Rate* is a binary variable equal to one if a student obtained a 5. A grade below 1.5 is identified as *Top Grade*. *Dropout Rate* is a binary variable equal to one if a student did not proceed to the third semester. *One-Year Break* is equal to 1 if a student enrolled earliest 18 month after high school graduation. *Academic HS* is equal to one if a student attended an academic high school. *Demographic and Study-Related Covariates* include in the regressions on the grades dummies for sex, nationality, majors, semesters, and exams. In the regressions on the outcomes *One-Year Break* and *Dropout Rate*, the covariates only comprise dummies for sex, nationality, and majors. Clustered standard errors at the student level are reported in parentheses.

Chapter 3

Accountability in Higher Education: The Impact of High-Stakes Testing on Academic Achievement

3.1 Introduction

About one third of the undergraduates in OECD countries do not finish their studies but drop out earlier (OECD, 2010). The issue is at the forefront of many political discussions as high drop-out rates especially in later semesters of a program are costly for all participating parties. Therefore, many European countries have implemented policies holding students more accountable for their academic achievement in order to reduce late drop-outs (for an overview, see Vossensteyn et al., 2015). One of the most common policies is high-stakes testing, implemented, among others, at Dutch, German, and Swiss universities. First-year students are not allowed to proceed with their studies unless they have successfully passed certain introductory courses. The aim of such a policy is to test at an early stage whether a student has the required skills to graduate, and to exclude those students who do not satisfy the requirements as early as possible. The effectiveness of such a policy, however, has been barely investigated so far.

Our paper helps to fill this gap by investigating the effect of a high-stakes testing policy that has been implemented in a large German federal state. In order to be allowed to continue their studies, undergraduates in Baden-Württemberg have to pass so-called orientation exams in the first two years. Students who fail one of these exams twice lose the eligibility to study their subject at every German university. Thus, orientation exams provide students with a strong incentive to put a lot of effort into their studies right at the beginning, and may reduce the inefficiency of late drop-outs. The success of this policy in achieving its goal, however, strongly depends on the implemented standards. If the requirements in an orientation exam are set too low (or too high), the policy will miss its goal to exclude (only) those students early who do not have the required skills to successfully complete their studies.

We investigate the effectiveness of this high-stakes testing policy in achieving its goal by using two quasi-experimental designs. In the first part of this paper, we use administrative panel data, aggregated at the study program level[1], on all university students in Germany from 1997 to 2003 to assess the average causal effect of high-stakes testing on drop-out within the first two years. As orientation exams were only introduced at universities in Baden-Württemberg, we can disentangle the effect of the policy from cohort and state fixed effects by estimating a difference-in-differences (DiD) model. This allows us to gain a first idea of the effectiveness of the policy in achieving its goal. In the second part of this paper, we apply a sharp regression discontinuity design (RDD) to administrative, student-level panel data of the Economics program at the University of Konstanz, one of nine universities in Baden-Württemberg. This enables us to investigate in more detail the

[1] A study program refers to one specific program, at one specific university.

mechanism of the high-stakes testing policy, as we observe students and their academic achievement up to graduation (or drop-out, respectively).

If the policy achieves its goal to exclude (only) those students who do not have the required skills to successfully complete their studies, we should find – in the first part of this paper – a significant increase in the average drop-out rate after two years. At the same time, we should not find – in the second part of this paper – an effect of narrowly failing the first attempt of an orientation exam on the probability to graduate, as we compare students of basically the same ability. Such findings would also be supported by the existing evidence on the effects of high-stakes testing policies on study success. Both Arnold (2015) and Tafreschi and Thiemann (2016) find that the introduction of academic dismissal policies in the Netherlands and Switzerland, respectively, increased first-year drop-out rates, but did not significantly affect graduation rates.[2]

Our DiD estimates show that the introduction of orientation exams in Baden-Württemberg increased average drop-out rates after two years by about three percentage points, which is equivalent to an increase of about 10 percent. The analysis in the second part of this paper reveals that freshmen students who have bad luck and narrowly fail an orientation exam are by about 16 to 19 percentage points more likely to drop out after the first semester. This effect diminishes over time but persists until the end of their studies, as we still find a 10 to 13 percentage points difference in the probability to graduate.

The paper proceeds as follows: In Section 3.2, we describe the development of high-stakes testing in higher education and present existing research findings. Section 3.3 covers our DiD analysis to identify the treatment effect of orientation exams on early drop-out. Section 3.4 presents our RD analysis which allows us to investigate the mechanism of orientation exams in more detail. Section 3.5 concludes.

3.2 Accountability in Higher Education

In the last ten to fifteen years, many European countries have taken measures to improve study success in higher education, though the policies differ conceptually (for an overview, see Vossensteyn et al., 2015). 'Curriculum design policies', for example, are thought to make students aware of their competences and to make them reflect on their program choice at an early phase of study. Prominent examples of such curriculum design policies are compulsory introductory study phases, or exams, at the beginning of a program. Students cannot continue their studies unless they have successfully completed all compulsory courses. Consequently, curriculum design policies aim to improve study success by holding

[2]Tafreschi and Thiemann (2016) find some positive effects on time to graduation, but this result is not robust and not significant across specifications.

students more accountable for their academic achievement, but also aim to force early drop-outs and switches.

The idea of holding students, but also teachers and institutions, more responsible for academic achievement originated in the US, where similar policies are implemented at the secondary school level since the 1980s. With the passage of the No Child Left Behind Act in 2002, schools and school districts must develop curriculum and achievement standards and measure student learning through tests that are aligned with the standards, while public funding of higher education institutions is often directly linked to the achievement of objective benchmarks (see Betts and Costrell, 2001; Conner and Rabovsky, 2011). A key feature of such standards-based accountability systems is that consequences are associated with how well students, teachers, or institutions progress toward their learning objectives. For high school students, for example, consequences may include tests that determine whether or not they graduate from high school, or are promoted to the next grade.

At the tertiary level, high-stakes testing is less common in the US so far, though the emphasis on accountability as an educational reform policy grew considerably over the last two decades (see, e.g., Alexander, 2000; Conner and Rabovsky, 2011). Many federal states have changed the way they finance higher education institutions. Until the mid 1990s, state funding mainly depended on the number of students enrolled. As student numbers increased, and as a consequence the variety in student qualifications, more and more states required higher education institutions to set and report performance measures to be eligible for continued state funding. By holding higher education institutions more accountable for students' study success, retention rates and degree production should be improved (Hillman et al., 2015). With similar intentions, the German federal state of North Rhine-Westphalia started in 2015 to pay universities 4000 euros for every graduate to create an incentive for lower drop-out rates especially in later years of a program. Other states plan to follow this approach.

3.2.1 Prior Research

In a theoretical paper, Betts and Costrell (2001) discuss the possible incentive effects of high-stakes testing on academic achievement, and how the effect depends on where the student is in the ability distribution. For students who are narrowly below the margin of passing, it should only take a small increase in their effort to pass the test. By comparing the costs, i.e., the effort, and the benefits of trying harder, i.e., passing the test, students at the margin of passing should respond to the incentive and work harder. In contrast, students who are far below or far above the margin have little incentives to change their effort.

Early studies examining the effect of high school graduation exams on academic

achievement find a positive relation (Bishop, 1998; Neill and Gayler, 2001). However, when using more extensive data and better controls for prior student achievement, Jacob (2001) finds no effect. Using data from Texas, Martorell (2004) finds that failing the high school exit exam lowers earnings, but only among individuals whose work experience indicates a strong attachment to the Texas labor market. Carnoy and Loeb (2002) and Hanushek and Raymond (2005) use state-level panel data to investigate the effect of accountability systems on academic achievement. Both studies find that students in high-accountability states achieved on average significantly greater gains on the National Assessment of Educational Progress (NAEP) tests. Investigating the effectiveness of an accountability policy implemented in the Chicago public schools in 1996-97, Jacob (2005) finds that math and reading achievement increased sharply following the introduction of the policy. His results further suggest that the observed achievement gains were driven by increases in test-specific skills and student effort, while teachers responded strategically to the incentives along a variety of dimensions - by increasing special education placements, preemptively retaining students, and substituting away from low-stakes subjects such as science and social studies. A related study by Jacob and Lefgren (2004) using a RD approach finds that the Chicago accountability program substantially increased academic achievement among third graders, but not among sixth graders.

Convincing quasi-experimental evidence for the effect of high-stakes testing on academic achievement at the tertiary level is scarce. Arnold (2015) investigates an academic dismissal policy that has been implemented gradually at many Dutch universities in the early 2000s, the so-called binding study advice (BSA). Students are dismissed from their university if they do not achieve a minimum number of credits after the first year. Estimating a fixed effects model, Arnold (2015) finds that the four-year completion rate is not significantly affected by the introduction of a BSA. However, his estimates are likely to suffer from endogeneity, as he cannot account for a potential selection of universities implementing the BSA. Tafreschi and Thiemann (2016) investigate the effect of a compulsory introductory study phase, implemented at the University of St. Gallen. Students who do not meet a certain performance requirement after the first year are forced to repeat all first year courses before they are allowed to move to the second year. Applying a RD design, Tafreschi and Thiemann (2016) find that students who have to repeat the first year are about 10 percentage points more likely to drop out of the program, but achieve half a standard deviation better grade point averages if they stay in the program.

Another strand of the literature shows that high-powered incentives may also lead to unwanted, strategic behavior. Holmström and Milgrom (1991) show that agents will focus on the most easily observable aspects of a multi-dimensional task if the incentive scheme is based on objective criteria. In a school context, teachers may shift resources away from

low-stakes subjects toward high-stakes ones, may neglect non-marginal students, or may teach to the test, while students may concentrate their study effort on high-stakes exams and adopt their learning strategies accordingly. Students who fail a high-stakes exam may further be discouraged and choose to drop out of school or higher education rather than to retake the exam in the future (Clark and Martorell, 2014).

Overall, research suggests that high-stakes testing can raise student effort and performance, depending on the context and type of policy implemented. However, research also shows that high-stakes testing can have severe consequences for those who fail to pass an high-stakes test, and students, teachers, and institutions may respond in unintended ways when faced with new incentive structures.

3.3 The Treatment Effect of Orientation Exams

In the following section, we investigate the average causal effect of the introduction of high-stakes exams in Baden-Württemberg on drop-out within the first two years. This allows us to gain a first idea of the effectiveness of the policy in achieving its goal.

3.3.1 Institutional Setup

In September 1999, the federal state of Baden-Württemberg passed a bill that obliged universities to introduce orientation exams in each study program from the winter term 2000/01 onward.[3] At this time, no other federal state had implemented a comparable high-stakes testing policy. We exploit this variation between states over time to identify the effect of the introduction of orientation exams on drop-out.

The aim of the legislation was to test at an earlier stage whether a student has the required skills to successfully proceed in their program, and to exclude those students who do no satisfy the requirements as early as possible (Bölke and Haug, 2009, p. 224). Before the passage of the bill – and in all other German federal states – the first high-stakes exams were written at the end of the second, or rather third year. The amendment now required students in Baden-Württemberg to take the first attempt of an orientation exam until the end of the first year. If students failed the exam, they had to pass the retake exam latest by the end of the third semester.[4] Otherwise, they were withdrawn from their studies and lost the eligibility to study their subject at every German university.

[3]Since the winter term 2014/15, it is up to the universities whether they implement orientation exams in their study programs.

[4]Exemptions were possible due to certified illness, or specific examination rules of a faculty (e.g., retake exams may only be taken one year after the first attempt, i.e., in the fourth semester).

3.3.2 Data

To investigate the effect of the introduction of orientation exams on the average drop-out rate of university students, we use administrative data on all higher education students who were enrolled in Germany between 1997/98 and 2003/04. The administrative data records every winter term (October to March) all students enrolled in a public university. In total, the data contain about 14.8 million student-semester observations. In addition to detailed information on the individual study situation such as the place of study, the study program, the enrollment year, and the semester, the data contain a limited number of student background variables such as gender, age and nationality.

As the data set provides no individual panel identifier, we aggregate the data at the study program level. The amendment only affected the state universities in Baden-Württemberg. Therefore, we exclude in a first step all non-state universities from our analysis. We further restrict the sample to students who were enrolled in their first study program as the incentive effects of the introduction of orientation exams may differ for students enrolled in their second or third study program. To obtain a balanced panel, we further exclude the universities from Hamburg in our baseline specifications as they were not sampled in the 1997/98 wave. Our final sample comprises about 3.75 million student-semester observations from 2421 study programs, 80 universities, and 15 federal states, which were collected between 1997/98 and 2003/04.

Our outcome variable, the drop-out rate, is defined in the following way: First, we determine the number of enrolled students in study program p at university u of cohort c. The drop-out rate among first-year students is then computed by calculating the difference between the number of first- and second-year students of study program p at university u of cohort c, and dividing it by the number of first-year students of study program p at university u of cohort c. We thus follow cohorts of first-year students and calculate drop-out rates for each cohort. In the same way, we compute the drop-out rates among second-year students of study program p at university u of cohort c. Because of the one-year time lag that is necessary to observe the second-year students, the 2001/02 cohort is the last cohort we consider in our analysis. In our baseline specification, we further restrict the analysis to drop-out rates greater or equal to zero.[5]

The fact that the data set provides no individual panel identifier implies that we cannot identify students who switched universities but stayed in the same study program, as the semester information is taken over by the new university. Thus, our outcome variable

[5]Because of real increases in student numbers over time within one cohort of one specific study program, but also because of measurement and coding errors in the data set, about six percent of the drop-out rates are in the negative domain. More than half of these negative drop-out rates reflect increases in student numbers of less than five students. We check the validity of our baseline estimations in the robustness section.

Table 3.1: Descriptive Statistics

	Treatment		Control	
	Pre	Post	Pre	Post
Males	0.495	0.458	0.451	0.444
	(0.238)	(0.232)	(0.242)	(0.242)
German	0.889	0.874	0.917	0.908
	(0.104)	(0.110)	(0.100)	(0.107)
Academic HS	0.751	0.743	0.782	0.782
	(0.123)	(0.130)	(0.145)	(0.142)
Full-Time Students	1	1	0.997	0.997
	(0.000)	(0.006)	(0.034)	(0.030)
No. of Study Programs	596	429	4885	3362

Note: Based on 1997-2001 university cohorts. The treatment group consists of university students from Baden-Württemberg, the control group of university students from all other German states except for Hamburg. The figures show the mean fractions of the respective group of the pre-reform cohorts, i.e., of the 1997 to 1999 cohorts, and the post-reform cohorts, i.e., of the 2000 to 2001 cohorts. Standard errors are reported in parenthesis.

captures both switches between universities within a study program, and withdrawals from a study program.

Table 3.1 presents summary statistics based on the students of the 1997/98 to 2001/02 cohorts. The first column reports mean characteristics of students from Baden-Württemberg of the 1997/98 to 1999/00 cohorts, the second column of students of the 2000/01 and 2001/02 cohorts. Mean characteristics of the respective cohorts from the control states are reported in columns three and four. Overall, Table 3.1 shows that the treatment and the control group have fairly similar characteristics, although the fraction of males, Germans, as well as academic high school graduates decreased slightly stronger in Baden-Württemberg than in the control states. However, when we estimate our baseline specification using the fraction of males, Germans, as well as academic high school graduates as the dependent variables, we find no statistically significantly different development between the treatment and the control group.

3.3.3 Empirical Strategy

To assess the average causal effect of the introduction of orientation exams on drop-out within the first two years, we exploit variation between states over time. Since the winter term 2000/01, students of all study programs who were enrolled at a university in Baden-Württemberg had to pass orientation exams until the end of their third semester, while students being enrolled at a university in another federal state had no such exams. This

allows us to isolate the effect of the introduction of orientation exams from cohort and state fixed effects by estimating the following difference-in-differences model using ordinary least squares (OLS):

$$Y_{pus}^{c} = \alpha\, BW_{pus} + \beta\, Cohort^{c} + \gamma\, Cohort^{c} \times BW_{pus} + \mu_s + \delta\, X_p + \eta_{pus}^{c} \tag{3.1}$$

where Y_{pus}^{c} denotes the drop-out rate in the respective time frame of cohort c of study program p at university u, located in state s. BW_{pus} represents a dummy variable equal to one if the university is located in Baden-Württemberg. $Cohort^{c}$ captures cohort fixed effects, and μ_s additional state fixed effects. X_p is a vector of demographic covariates comprising the share of male students, the share of students with German citizenship, and the number of freshmen by cohort and state. Error terms η_{pus}^{c} are clustered at the study program level to account for the presence of heteroscedasticity.

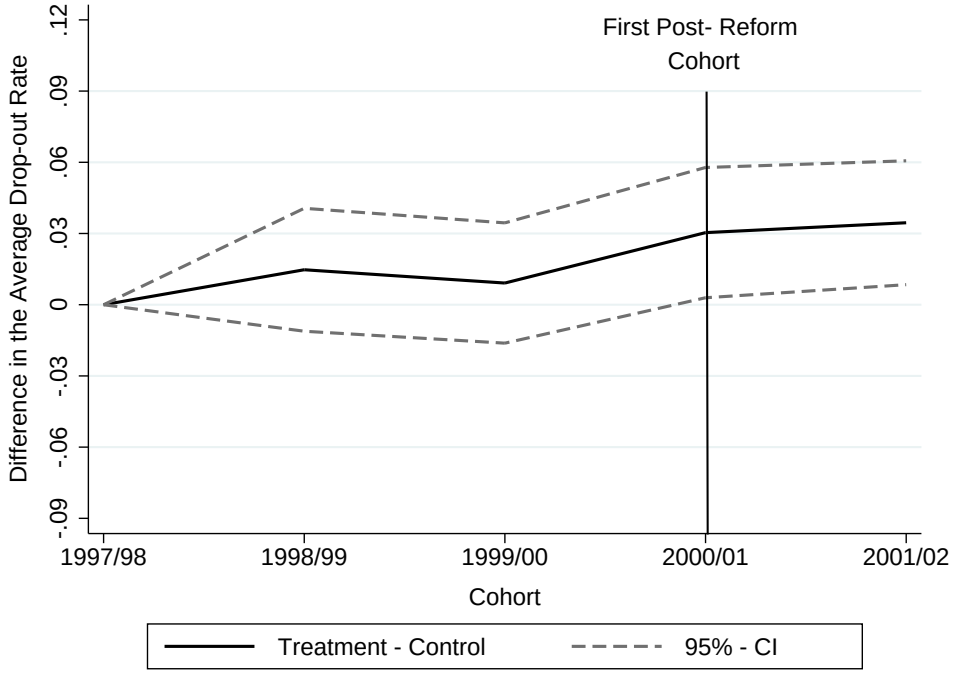
The coefficient of interest is γ, which is the average treatment effect on the treated of the introduction of orientation exams on the drop-out rate of university students.[6] It measures the change in the drop-out rate after the passage of the bill, relative to before the passage of the bill, in study programs in Baden-Württemberg relative to study programs in the control states. The key identifying assumption is that there were no further changes at the program level parallel to the amendment affecting the drop-out rates, or the composition of the student bodies at study programs in Baden-Württemberg or the control states. Thus, we assume that the underlying trends in the drop-out rate would have been the same for both the treatment and the control group in the absence of the amendment.

Figure 3.1 presents suggestive evidence for the validity of the common trend assumption (CTA). The figure shows the difference in the drop-out rate after two years between students enrolled in a study program at a university in Baden-Württemberg and students enrolled in a study program at a university in the control states over time. The dashed lines represent 95% confidence intervals.[7] The solid, vertical line indicates the first treatment cohort. The figure suggests that the CTA is fulfilled for Equation (3.1) as the difference in the drop-out rate between the treatment and the control group is never statistically significantly different from zero for the pre-reform cohorts. With the 2000/01 cohort this changed: The drop-out rate after two years at universities in Baden-Württemberg rose significantly relative to the drop-out rate at universities in the control states for both the 2000/01 and the 2001/02 cohort. This is a first hint that the introduction of orientation

[6] As the reform was barely mentioned in the media, and as we do not find evidence for selective enrollment in universities due to the reform (see Figures 3.2 and 3.3), the reported effect should be identical to the average treatment effect.

[7] The data are obtained by OLS regressions of Equation (3.1) without covariates using our baseline sample.

Figure 3.1: Difference in the Average Drop-out Rate After the Second Year Between University Students from Baden-Württemberg (Treatment Group) and University Students from All Other German States (Control Group) by University Cohort



Note: The solid line represents the difference in the average drop-out rate, calculated at the study program level, after the second year between treatment and control students. The dashed-lines represent 95% confidence bands. The vertical line indicates the treatment year. The treatment group consists of university students from Baden-Württemberg, the control group of university students from all other German states except for Hamburg.

exams forced significantly more early drop-outs.

Our DiD analysis may still provide misleading estimates because of an omitted variables bias arising from unobserved and uncontrolled differences between the treatment and the control group. In particular, Equation (3.1) will provide a biased estimate of the causal effect of orientation exams on average drop-out rates if high school graduates opted less often for universities in Baden-Württemberg in order to evade the reform. Figures 3.2 and 3.3, however, do not confirm this conjecture: The number of freshmen students at universities in Baden-Württemberg follows the same trend as the number of freshmen students at universities in the control states, and the same holds for the evolution of freshmen students at universities and universities of applied sciences in Baden-Württemberg.[8] If anything, the fraction of students who enrolled at a university in

[8]When we estimate Equation (3.1) using the number of freshmen students as the dependent variable, we find no evidence for a statistically significantly different enrollment pattern at universities in Baden-Württemberg and universities in the control states. When we only consider the freshmen students in Baden-Württemberg and replace in our baseline specification the dummy variable for Baden-Württemberg with a dummy variable for universities, we also find no evidence for a different enrollment pattern of

Figure 3.2: Number of Freshmen Students by Cohort, Baden-Württemberg vs. Control States

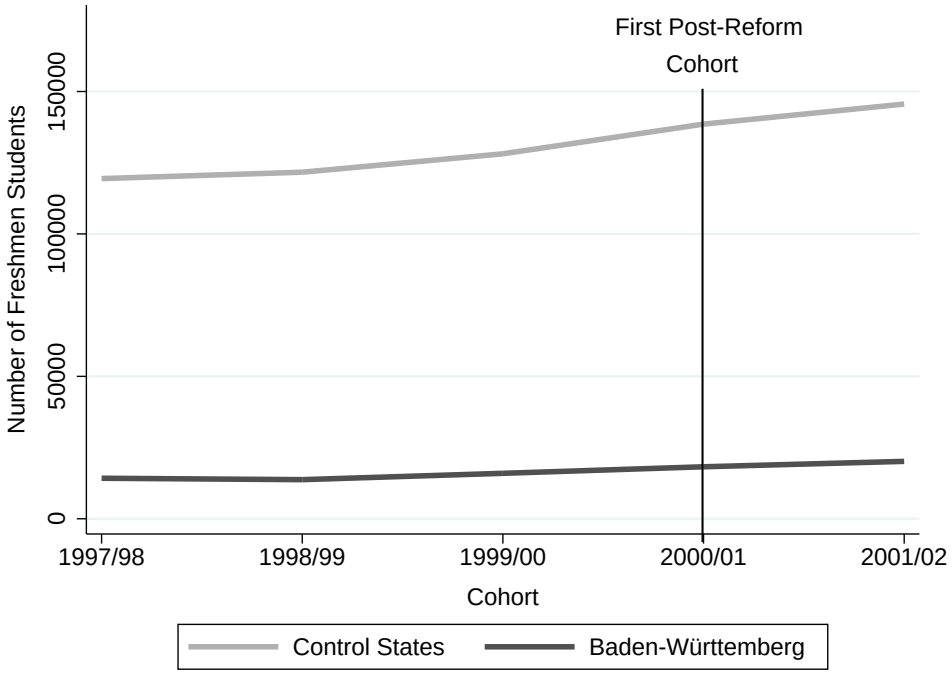
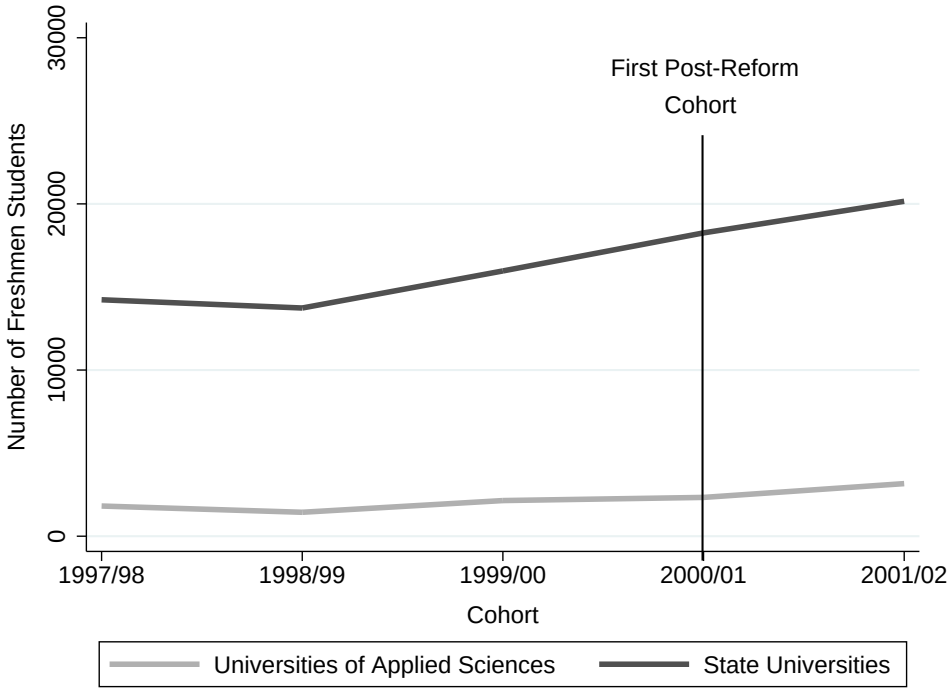


Figure 3.3: Number of Freshmen Students by Cohort, Universities vs. Universities of Applied Sciences in Baden-Württemberg



university and university of applied sciences students.

Baden-Württemberg slightly increased relative to the fraction of students who enrolled at a university of applied sciences.

3.3.4 Results

In the following, we present DiD estimates of the effect of the introduction of orientation exams on average drop-out rates within the first two years. We start with the results of our baseline specification. Afterwards, we examine the sensitivity of our baseline specification to changes in the sample restrictions and the model specifications. Having demonstrated the robustness of our estimates, we investigate whether the effects vary across subgroups.

The Treatment Effect of Orientation Exams on Early Drop-out

Table 3.2 presents the results of the DiD estimates of Equation (3.1). Columns one, four, and seven report the effect of orientation exams on the respective drop-out rate without any controls. The estimates in columns two, five, and eight include additional state fixed effects. The DiD estimates of our baseline specification are reported in columns three, six, and nine, and additionally include a set of demographic controls. The regressions are based on the students of the 1997/98 to 2000/01 cohorts. The control group consists of the university students from all other German states except for Hamburg.[9]. As depicted in Figure 3.1, we do not find evidence for a violation of the CTA when estimating the effect of the introduction of orientation exams on the first-year, second-year, or two-years drop-out rate with a full set of interaction terms. The inclusion of additional state fixed effects as well as the set of further controls does also barely affect our estimates, supporting the assumption that the composition of the treatment and the control group is random.

Table 3.2 reveals a statistically and economically significant effect of the introduction of orientation exams on the likelihood to drop out within the first two years. This effect is solely driven by the increase in the second-year drop-out rate: Students of the 2000/01 cohort in Baden-Württemberg who were the first to take orientation exams were about three percentage points more likely to drop-out of their study program in the second year than students of earlier cohorts. This is equivalent to an increase in the second-year drop-out rate of about 10 percent. Thus, our results suggest that the orientation exams were implemented properly, i.e., according to the regulations of the amendment: Students who failed the first attempt of an orientation exam, had to take the retake exam in the second year, and were dismissed from their study program if they failed again.

[9]Hamburg was not sampled in the 1997/98 wave. Therefore, we exclude Hamburg from our baseline specification to obtain a balanced panel. Including Hamburg, however, does not significantly change the estimates (see Table 3.4, column six).

Table 3.2: DiD Estimates: The Effect of the Reform on Drop-out

	First- and Second-Year			First-Year Only			Second-Year Only		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reform Effect									
BW x 2000	0.030**	0.028**	0.025*	0.007	0.006	0.002	0.034***	0.033***	0.032***
	(0.014)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.011)	(0.012)	(0.012)
Pre-Reform Trend									
BW x 1999	0.009	0.006	0.005	-0.003	-0.004	-0.005	0.020*	0.018	0.018
	(0.013)	(0.013)	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
BW x 1998	0.015	0.011	0.008	0.005	0.003	-0.001	0.017	0.016	0.015
	(0.013)	(0.013)	(0.013)	(0.011)	(0.011)	(0.011)	(0.013)	(0.013)	(0.013)
Cohort Fixed Effects									
2000	0.002	0.003	-0.001	0.010**	0.011**	0.007*	-0.007*	-0.007*	-0.009**
	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
1999	0.005	0.006	0.003	0.012***	0.013***	0.010**	-0.007*	-0.006	-0.008*
	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
1998	0.003	0.004	0.002	0.006	0.006	0.005	0.001	-0.000	-0.001
	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Baden-Württemberg Fixed Effect	0.010	-0.011	-0.007	0.015	0.004	0.009	-0.005	-0.018	-0.015
	(0.013)	(0.050)	(0.049)	(0.011)	(0.048)	(0.046)	(0.010)	(0.030)	(0.030)
Additional State Fixed Effects	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Further Controls	No	No	Yes	No	No	Yes	No	No	Yes
No. of Observations	7388	7388	7388	7388	7388	7388	7388	7388	7388

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

Note: Linear probability models. Regressions are based on 1997-2000 university cohorts. The dependent variable is the drop-out rate at the end of the respective year, calculated at the program level. The treatment group consists of university students from Baden-Württemberg, the control group of university students from all other German states except for Hamburg. *Further Controls* include the share of male students, the share of students with German citizenship, and the number of freshmen by cohort and state. Standard errors, reported in parentheses, are clustered at the study program level.

Robustness Checks

In the following, we show that the results presented above are robust to a variety of alternative specification choices and validity checks. In particular, we present estimates of Equation (3.1) when we relax the domain restriction concerning our outcome variables, investigate the effect on the second cohort that was affected by the amendment, perform placebo tests, and check whether our results differ when we vary the sample, i.e., when we include Hamburg or restrict the sample to full-time students.

In our baseline specification, we only consider drop-out rates that are greater or equal to zero. However, because of real increases in student numbers over time within one cohort of one specific study program which may occur because of students switching study programs, but also because of measurement and coding errors in the data set, about six percent of the drop-out rates are in the negative domain.[10] Columns two to four of Table 3.3 present estimates of Equation (3.1) when we relax the domain restriction concerning our outcome variables, and allow for increases in cohort sizes over time of up to 20 percent. Table 3.3 shows that our baseline results are very robust to these alternative domain restrictions, as both the point estimates and the standard errors do barely change across specifications.

Since we cannot determine all drop-out measures for the last two cohorts in our data set, we focus on the students of the 1997/98 to 2000/01 cohorts in our baseline regressions. However, the effects should be similar for later cohorts. Column two of Table 3.4 presents the estimates for the second cohort affected by the amendment, i.e., the Baden-Württemberg students of the 2001/02 cohort. The control group consists of university students from all other German states except for Hamburg. The regressions are based on the 1997/98 to 2001/02 cohorts where the 2000/01 cohort is excluded from the analysis. The results are very similar to our baseline results: While the point estimate for the likelihood to drop out in the second year is identical to the estimate of the 2000/01 cohort, the point estimate for the likelihood to drop out in the first year is slightly larger, yielding a slightly higher drop-out rate after two years. Overall, the results of the second cohort strongly support our baseline DiD results.

If the effects are causal effects of the introduction of orientation exams, the effects should be present for Baden-Württemberg students of the 2000/01 cohort and later with no corresponding significant effects for students of earlier cohorts, or from other states. In columns three to five of Table 3.4 the results of the respective placebo tests are shown. While columns three and four present estimates of placebo tests in the timing of the reform, column five presents estimates of a placebo test where we assign the treatment status to Bavarian students, i.e., to students from a large neighboring state of Baden-

[10] More than half of these negative observations reflect increases of less than five students.

Table 3.3: DiD Estimates: Robustness Checks

	Baseline	$\geq -5\%$	$\geq -10\%$	$\geq -20\%$
Drop-out, 1st and 2nd Year	0.025*	0.026*	0.022	0.022
	(0.014)	(0.014)	(0.014)	(0.014)
No. of Observations	7388	7691	8017	8343
- 1st Year Only	0.002	-0.001	-0.006	-0.002
	(0.013)	(0.013)	(0.013)	(0.012)
No. of Observations	7388	7691	8017	8343
- 2nd Year Only	0.032***	0.036***	0.038***	0.033***
	(0.012)	(0.011)	(0.011)	(0.012)
No. of Observations	7388	7691	8017	8343
Cohort Fixed Effects	Yes	Yes	Yes	Yes
Baden-Württemberg Fixed Effect	Yes	Yes	Yes	Yes
Additional State Fixed Effects	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes

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

Note: Linear probability models. Regressions are based on 1997-2000 university cohorts. The dependent variable is the drop-out rate at the end of the respective year, calculated at the study program level. The treatment group consists of university students from Baden-Württemberg, the control group of university students from all other German states except for Hamburg. *Further Controls* include the share of male students, the share of students with German citizenship, and the number of freshmen by cohort and state. Standard errors, reported in parentheses, are clustered at the study program level.

Württemberg which is also comparable regarding the education system and infrastructure. The placebo treatment group consists in column three of the students of the 1999/00 cohort of Baden-Württemberg, and in column four of the students of the 1998/99 cohort of Baden-Württemberg. The control group consists of university students of the respective cohorts from all other German states, except for Hamburg, and except for Baden-Württemberg in column five. The results show that there is no significant difference in the drop-out rates between treatment and control students in the pre-treatment years. There is also no significant difference between Bavarian students and students from the remaining control states.[11] Thus, the placebo tests support a causal interpretation of our baseline results.

To obtain a balanced panel, we exclude the universities of Hamburg from our baseline specification as Hamburg was not sampled in the 1997/98 wave. A conjecture may be that the effects differ if we include the universities of Hamburg in our sample. Column six of Table 3.4 presents the effect of the introduction of orientation exams on the first treatment cohort including Hamburg in the control group. The results are almost identical to our baseline results, showing that our baseline specification is robust to the inclusion of

[11] The same results are obtained when assigning the treatment status to another state than Bavaria.

Table 3.4: DiD Estimates: Robustness Checks, Cont'd

	Baseline	Second Cohort	Placebo - 1 Year	Placebo - 2 Years	Placebo Bavaria	Including Hamburg	Full-Time Students
Drop-out, 1st and 2nd Year	0.025*	0.031**	-0.003	0.011	-0.021	0.025*	0.025*
	(0.014)	(0.014)	(0.012)	(0.014)	(0.013)	(0.014)	(0.014)
No. of Observations	7388	7365	5481	3635	6585	7587	7386
- 1st Year Only	0.002	0.009	-0.007	0.000	-0.016	0.003	0.002
	(0.013)	(0.012)	(0.011)	(0.012)	(0.011)	(0.013)	(0.012)
No. of Observations	7388	7365	5481	3635	6585	7587	7386
- 2nd Year Only	0.032***	0.032**	0.008	0.016	-0.007	0.032***	0.032***
	(0.012)	(0.013)	(0.011)	(0.013)	(0.011)	(0.012)	(0.012)
No. of Observations	7388	7365	5481	3635	6585	7587	7386
Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Baden-Württemberg Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Additional State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Note: Linear probability models. The dependent variable is the drop-out rate at the end of the respective year, calculated at the study program level. The treatment group consists of university students from Baden-Württemberg, the control group of university students from all other German states except for Hamburg (unless stated otherwise). *Further Controls* include the share of male students, the share of students with German citizenship, and the number of freshmen by cohort and state. Standard errors, reported in parentheses, are clustered at the study program level.

the universities of Hamburg.

A similar concern may be that our baseline results are to a large extent explained by the effect on the part-time students who are potentially hit more strongly by the introduction of orientation exams, and are consequently more likely to drop out of university (voluntarily) in case of failure. Column seven of Table 3.4 shows the estimates of Equation (3.1) when we restrict the sample to full-time students. The point estimates are identical to the point estimates of our baseline specification, showing that our baseline DiD results are not explained by the effect on part-time students.

Subgroup Analysis

The analysis thus far has focused on the average causal effect of the introduction of orientation exams on university students' probability to drop out within the first two years, finding a significant increase in the second-year drop-out rate. Additionally, there could be important heterogeneity in the treatment effect across subgroups. For example, the effect may be less pronounced or non-existent in study programs with highly able students as these students are likely to pass an exam anyway, while the effect may be more pronounced in study programs with a rather average student body in terms of ability. In the following, we present estimates of our baseline specification when we perform the analysis for the subgroup of general medicine students, i.e., a group of high ability students.[12] As we are using data of the Economics Department of the University of Konstanz in the second part of this paper, we additionally perform the DiD analysis for the subgroup of economics students, also to provide further evidence for the external validity of our RDD estimates.

The first three columns of Table 3.5 present the estimates of Equation (3.1) for the subgroup of medical students; the last three columns present the estimates for the subgroup of economics students. The estimates of our baseline specification are reported in columns three and six, respectively, and include cohort and state fixed effects as well as a set of demographic controls. Given that grading standards in medicine and economics were not altered differently with the introduction of the orientation exams, our results show that there is significant heterogeneity with respect to ability. While we find no significant effect in medicine, the drop-out rate after two years in economics has increased significantly. Considering the drop-out rates after the first and the second year separately, the drop-out rate of medical students after the first year even slightly decreased, although the effect is only statistically significant at the ten percent level. Overall, our results suggest that the effect of orientation exams on drop-out depends on the student body of a study program. While we do not find an effect for a study program where students are strongly preselected

[12]To be admitted to a study program in general medicine in Germany, a high school GPA of at least 2.1 was needed in 2000/01, where the high school GPA can range from 1 to 4 with 1 being the highest GPA and 4 being the lowest.

Table 3.5: DiD Estimates: Subgroup Analysis

	Medical Students			Economics Students		
	(1)	(2)	(3)	(4)	(5)	(6)
Drop-out, 1st and 2nd Year	0.040	0.032	0.027	0.117**	0.097**	0.088**
	(0.030)	(0.033)	(0.035)	(0.045)	(0.039)	(0.039)
No. of Observations	72	72	72	141	141	141
- 1st Year Only	-0.025	-0.026*	-0.025*	0.089	0.072	0.068
	(0.015)	(0.014)	(0.014)	(0.064)	(0.057)	(0.055)
No. of Observations	72	72	72	141	141	141
- 2nd Year Only	0.066**	0.058*	0.053	0.099	0.089	0.079
	(0.029)	(0.033)	(0.035)	(0.065)	(0.071)	(0.070)
No. of Observations	72	72	72	141	141	141
Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Baden-Württemberg Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Additional State Fixed Effects	No	Yes	Yes	No	Yes	Yes
Further Controls	No	No	Yes	No	No	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Note: Linear probability models. Regressions are based on 1997-2000 university cohorts. The dependent variable is the drop-out rate at the end of the respective year, calculated at the study program level. The treatment group consists of university students from Baden-Württemberg, the control group of university students from all other German states except for Hamburg. *Further Controls* include the share of male students, the share of students with German citizenship, and the number of freshmen by cohort and state. Standard errors, reported in parentheses, are clustered at the study program level.

in terms of high school performance, we do find a strong and significant effect for a study program with students of rather average ability.

3.4 Investigation of the Mechanism

The previous section has shown that the introduction of orientation exams significantly increased students' probability to drop out within the first two years. However, the DiD estimates do not allow us to draw a conclusion concerning the effectiveness of orientation exams in reducing late drop-outs for two reasons: First, we only observe drop-out rates until the end of the second year. The evolution of drop-outs after the second year remains unknown.[13] Second, the data set includes no individual panel identifier. Therefore, we could not differentiate between dropout and graduation, even if we would extend the data set beyond the winter term 2003/04.

Applying a sharp regression discontinuity design (RDD) to administrative, student-level panel data of the University of Konstanz, one of nine universities in Baden-Württemberg, allows us to shed light on the effectiveness as well as the mechanism of orientation exams. Exploiting exogenous variation in failing the first attempt of an orientation exam, we identify the effect of high-stakes testing on drop-out and graduation rates of marginal students, i.e., we compare students who just by chance passed the first attempt with students who just by chance failed the first attempt.

The main threat to the identification of the effect of high-stakes testing on subsequent university outcomes is selection into the treatment and the control group. In an ideal set-up, students are randomly assigned to the treatment and the control group. In fact, however, students who have to repeat an orientation exam are likely to differ significantly from students who do not have to repeat the orientation exam. Characteristics of the students, such as their ability, their motivation, or their parental background, will simultaneously affect treatment status and academic achievement, thereby leading to a biased estimate of the treatment effect. We overcome this identification problem by applying a sharp regression discontinuity design to administrative, student-level panel data of the Economics program at the University of Konstanz.

3.4.1 Institutional Setup

Similar to most Bachelor degree programs in Europe, the Economics program at the University of Konstanz consists of 180 ECTS credits with a workload of 60 ECTS credits per academic year. During the first three semesters, economics students at the University of

[13] The average length of tertiary studies were around five years at this time.

Konstanz have a set of compulsory courses, consisting of courses in business administration, economics and mathematics, which are tested during central examination periods. The main examination period is right after the lecture period, while the retake exams take place at the end of the semester break before the next semester starts. The only course assessment to obtain the required ECTS credits for each course is to pass a written exam after the lecture period.

Among the compulsory courses, economics students have to write two orientation exams in the first semester ('Mathematics I' and 'Introduction to Economics'). Students are automatically registered for the orientation exams in the first examination period, while they can choose the examination period for their other courses. Those who fail an Orientation exam at the first attempt are forced to repeat the exam at the next possible date. By failing twice, they immediately lose the eligibility to continue the program. Therefore, students have a strong incentive to increase their effort after having failed an orientation exam once.

Following a set of further compulsory exams in the second and third semester, students choose one out of six major tracks. In contrast to the curriculum of the first three semesters, the curriculum in the second half of the program is track specific. Students graduate on time if they complete the program within six semesters.

3.4.2 Data

We use unique, anonymized student-level panel data provided by the registrar's office of the University of Konstanz. The data set contains information on enrollment and academic achievement, i.e., grades and the exact number of points achieved in each orientation exam, for all economics students who started their undergraduate program in the winter term 2007/2008, or later. In total, the data comprise about 2,700 students, adding up to approximately 47,000 single student-exam observations. In addition, we have information on students' major, course of study, and the exact examination dates, i.e., the term and the examination period, of all attended courses until the summer term 2016. Furthermore, the data contains the following background characteristics: students' overall high-school grade point average (HSGPA), the type of high-school a student attended, the place and date of issue of the high-school diploma, students' year and month of birth, gender and nationality.

We impose the following restrictions on our sample: First, we only consider economics undergraduates who started the program between the winter term 2007/2008 and the winter term 2012/2013. Consequently, we restrict our sample to students who should finish their bachelor program until the summer term 2015.[14] Since the last semester for which

[14]The regular time schedule of the program is six semesters, but the average duration is about seven

we have data is the summer term 2016, we observe all students for at least eight semesters. Second, we only consider students who attended at least one of the orientation exams and did not drop out without attending one of them. Having imposed these restrictions, we end up with about 2,050 students and 37,000 exam observations.

In this second part of the paper, we investigate mainly two outcomes: First, whether a student drops out before finishing the program, and second, whether he graduates. Concerning the drop-out decision, we distinguish between drop-out after the first semester and drop-out after the third semester. Drop-out after the first semester captures mainly those students who do not pass the orientation exam and are therefore withdrawn from the program. Drop-out after the third semester coincides with the end of the basic studies, when students have to opt for a specific track afterwards, but also captures the drop-out of those students who – in the first semester – stayed absent from the second attempt with a medical certificate and failed to pass the orientation exam later on, in the third semester. Concerning graduation, we have detailed administrative information on the graduation status of each student. By observing the drop-out rate after the first and the third semester as well as the graduation rate allows us to observe both the short- and the long-term effect of failing an orientation exam on academic achievement, and thus, a crucial determinant of later labor market success (Altonji et al., 2012).

To create our assignment variable, we use the number of points achieved in the first semester at the first attempt of an orientation exam. In combination with the information on the exam date, we are able to identify in a first step those students who did not attend the exam due to illness or other reasons while being registered for it. Second, we are able to observe precisely how close a student was to the pass/fail cutoff in each exam. Our assignment variable is therefore the deviation from the cutoff, relative to the maximum number of points, in Mathematics I and Introduction to Economics, respectively, whereas we use the deviation from the pass/fail cutoff in Mathematics I in our baseline specification. In the robustness section, we perform the same analyses using the deviation from the cutoff in Introduction to Economics. The remaining description in this section, however, refers to our baseline specification.

Our control group consists of the students who did not fail the Mathematics I exam at the first attempt, whereas the treatment group consists of the students who did fail the first attempt. Thus, the students of the treatment group were forced to spend additional time on the preparation for the retake exam.[15]. Furthermore, they had to deal with the frustration of having failed an exam and feeling the pressure of ultimate failure, as the second attempt of an orientation exam is their last chance to pass before being withdrawn

semesters.

[15]The relevant lecture notes for an exam do not differ between the first and the second attempt. Thus, students sitting the retake exam will basically repeat practicing the lecture notes.

Table 3.6: Descriptive Statistics: Background Characteristics

	Full Sample	Failed	Passed	P-Value
German	0.936	0.926	0.939	0.41
	(0.006)	(0.014)	(0.007)	
Male	0.564	0.640	0.540	0.00
	(0.013)	(0.026)	(0.015)	
Enrolment Age	20.698	20.847	20.651	0.01
	(0.033)	(0.064)	(0.038)	
HSGPA	2.370	2.724	2.258	0.00
	(0.015)	(0.026)	(0.016)	
Academic HS	0.75	0.668	0.775	0.00
	(0.012)	(0.027)	(0.013)	
School Duration	13.174	13.365	13.114	0.00
	(0.021)	(0046)	(0.022)	
No. of Students	1478	1125	353	

Note: Descriptive statistics of students' background characteristics using the full sample, the students who failed Mathematics I at the first attempt, and the students who passed Mathematics I at the first attempt. P-values of t-tests for mean differences in the background characteristics between the two groups are presented in the very right column. Standard errors are reported in parentheses.

from the program.

Table 3.6 shows descriptive statistics of background characteristics for the full sample, the students who failed Mathematics I at the first attempt, and the students who passed Mathematics I at the first attempt. In the full sample, the average enrollment age is below 21 years. There are few more men than women enrolled in the program and the large majority of students is German (more than 90 percent). The average HSGPA is at 2.37. Considering the split samples, Table 3.6 reveals significant differences between the two groups in terms of their background characteristics. Students who passed Mathematics I at the first attempt graduated significantly earlier from high-school, achieved a significantly higher HSGPA of half a grade on average, and were significantly younger at the time of their enrollment. Among the students who failed Mathematics I at the first attempt, there are also significantly less men than women.

Considering the outcome variables, Table 3.7 reports significant differences in terms of drop-out and graduation rates between the students who passed, and the students who failed Mathematics I at the first attempt. While the drop-out probability of students who passed the exam are only at around four percent at the end of the first semester, the drop-out probability of students who failed are at about 40 percent. These figures rise to 13 and 60 percent, respectively, until the end of the third semester, and persists until graduation: While almost 80 percent of the students who passed Mathematics I at the

Table 3.7: Descriptive Statistics: Outcomes

	Full	Failed	Passed	P-Value
Drop-out I	0.123	0.402	0.044	0.00
	(0.009)	(0.026)	(0.006)	
Drop-out III	0.244	0.592	0.134	0.00
	(0.011)	(0.026)	(0.010)	
Graduation	0.662	0.266	0.787	0.00
	(0.012)	(0.024)	(0.012)	
No. of Students	1478	353	1125	

Note: Descriptive statistics of student outcomes using the full sample, the students who failed Mathematics I at the first attempt, and the students who passed Mathematics I at the first attempt. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. *Graduation* is a dummy variable equal to one if a student graduated. P-values of t-tests for mean differences in the outcome variables between the two groups are presented in the very right column. Standard errors are reported in parentheses.

first attempt graduate, only about 27 percent of the students who failed Mathematics I graduate.

3.4.3 Empirical Strategy

The main threat to the identification of the effect of failing an orientation exam on academic achievement is that students self-select into the treatment and the control group. The significant differences in the background characteristics shown in Table 3.6 support this assumption. Students being exposed to a high-stakes exam will not only systematically differ from students not being exposed to a high-stakes exam in terms of observable characteristics, but also in terms of unobservable characteristics, such as ability, motivation, or parental background. Therefore, a simple mean comparison of the outcomes of treatment and control students will provide only a very naïve estimate of the treatment effect (Angrist and Pischke, 2009, p. 196-203). To identify the true impact of failing an orientation exam on subsequent academic achievement, we exploit the discontinuous relationship between the number of points achieved in an exam and the probability to fail. This discontinuity generates plausible exogenous variation and allows us to identify the causal effect of failing an orientation exam on marginal students' academic achievement.

Because of the institutional regulations, all students who fail an orientation exam are automatically registered for the retake exam which is written at the end of the semester break. Observing the exact number of points achieved and having no non-complier in our data, we estimate a sharp RDD which can be described by the following equation (see

Angrist and Pischke, 2009, p. 253):

$$Y_{ic} = \alpha + \beta\, failed_i + \gamma\, f(deviation_i) + \delta\, X_i + \nu_c + \eta_{ic} \tag{3.2}$$

where Y_{ic} denotes the outcome of interest, i.e., either drop-out or graduation of student i of cohort c.[16] $failed_i$ is a dummy variable equal to one if a student failed the first attempt of an orientation exam. The assignment variable $deviation_i$ is the deviation in points achieved from the pass/fail cutoff and is included in our parametric specifications by a n^{th} degree polynomial function. X_i is a vector of demographic covariates comprising gender, age, nationality, HSGPA, and high school type attended. In all parametric specifications, we control for cohort fixed effects ν_c to account for time-varying conditions such as course contents and grading schemes. Error terms η_{ic} are clustered at the level of the assignment variable.

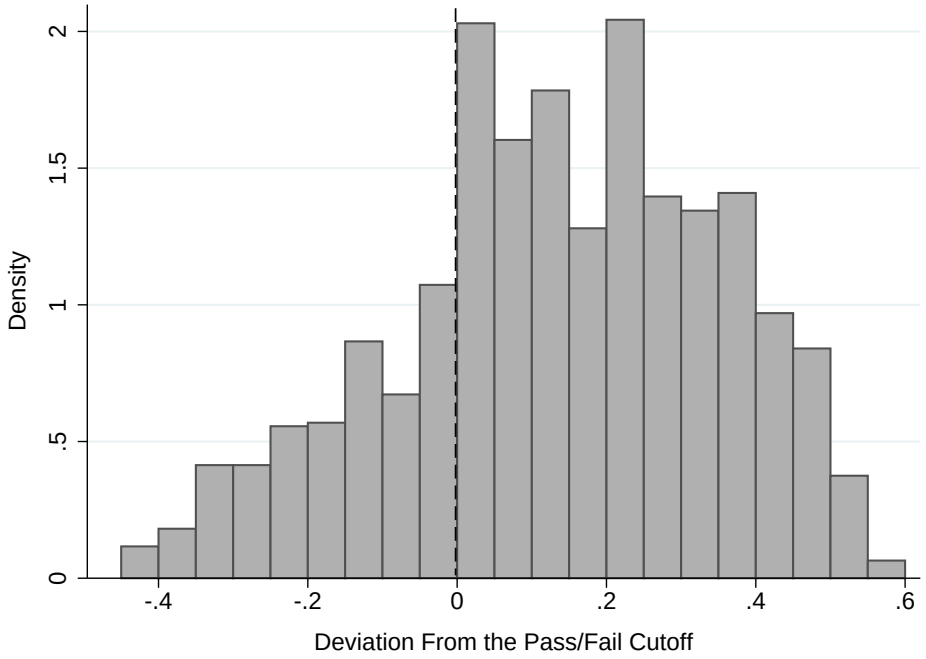
In our baseline parametric specification, we use symmetric 1st-, 2nd-, 3rd-, and 4th-degree polynomial functions for varying windows around the cutoff. Using asymmetric polynomials instead to account for different slopes left and right to the cutoff does not significantly change the results. For all outcomes, we use linear probability models, but results are robust to alternative specifications (e.g., Probit or Logit models). We also provide nonparametric estimates, relying on local linear regressions (Imbens and Lemieux, 2008). The optimal bandwidth is derived by cross-validation, and standard errors are computed using the bootstrap method as suggested by Imbens and Kalyanaraman (2012).

The coefficient of interest is β, which is the effect of failing the first attempt of an orientation exam on marginal students' subsequent academic achievement. The key identifying assumption underlying our framework is the local continuity assumption, which implies that the function $f(.)$ is continuous through the pass/fail cutoff and that characteristics change smoothly around the cutoff. Thus, students who are closely above the cutoff should be similar to those closely below the cutoff and would, on average, achieve the same outcomes if they had been exposed to the treatment. If the local continuity assumption holds, the estimated effect can be interpreted as the causal effect of failing the first attempt of an orientation exam on subsequent academic achievement (Imbens and Lemieux, 2008).

A common concern regarding the local continuity assumption is the possibility of precise manipulation of the assignment variable around the cutoff (Urquiola and Verhoogen, 2009). In our setting, local continuity is likely to hold because, on the one hand, students do not have precise control over the exact number of points they achieve in an exam, and on the other hand, exams are written anonymously such that a systematic manipulation of the

[16] Both outcome variables are dummy variables equal to one if student i dropped out or graduated from the program.

Figure 3.4: Histogram of the Assignment Variable



Note: The assignment variable is the deviation from the pass/fail cutoff, relative to the maximum number of points, obtained in Mathematics I at the first attempt.

cutoff from the evaluators is not possible.[17] Nevertheless, professors (or teaching assistants) have the possibility to move students from just below the cutoff to just above the cutoff. Figure 3.4 suggests that such a manipulation indeed happens: There are significantly more students just above the cutoff than just below the cutoff, resulting in an unsmooth trend of our assignment variable. For the validity of our design, this is problematic if professors move students with specific characteristics to one side of the cutoff. Intuitively, this is hardly possible as exams are written anonymously, and grading takes place without knowing whose exam is graded.

To test for non-continuity at the cutoff, we perform a McCrary test (McCrary, 2008). The test does not reject the null hypothesis, i.e., the continuity of the density of our assignment variable at the cutoff. We further estimate our baseline specifications using the background characteristics as the dependent variables to test for systematic sorting around the threshold. Table 3.8 presents the coefficients and standard errors using different windows around the cutoff. For none of the background characteristics, we find evidence for a discontinuity in any specification. Particularly important for our design is that students do not differ with respect to their HSGPA, the best proxy variable for study success (see e.g., Geiser and Santelices, 2007; Rothstein, 2004; Trapmann et al., 2007). However, as the

[17] Gerard et al. (2015) show that even in the case of a manipulated running variable, RD estimates are still informative in the sense that they partially identify the value of interesting causal parameters, while sorting around RD cutoffs is innocuous when manipulated units are similar to those unaffected by sorting.

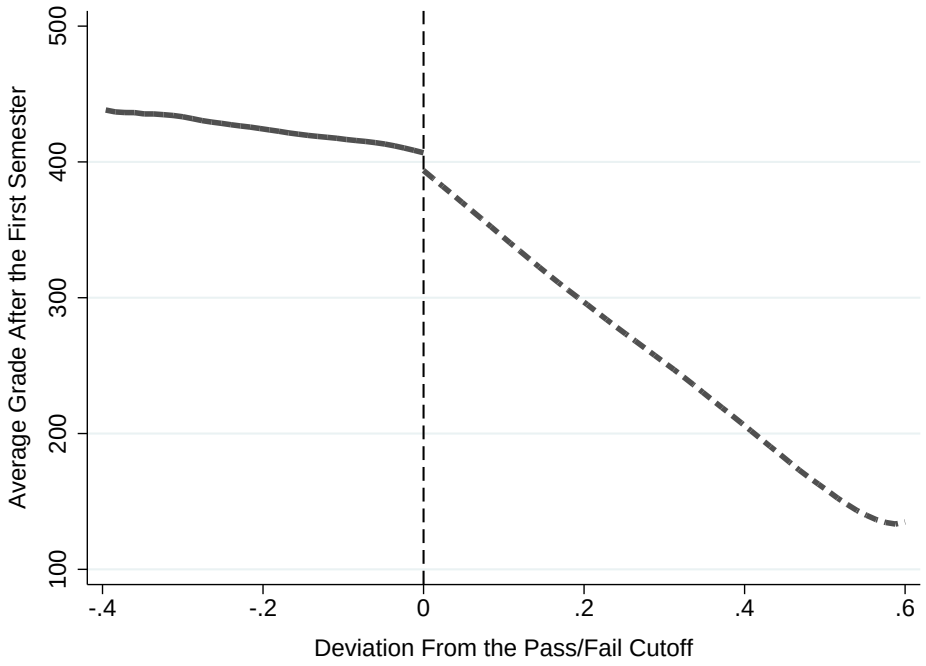
Table 3.8: RDD Estimates: Background Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
German	-0.00281	0.0349	0.0270	0.0432	0.0412	0.0393	0.0443	0.0369
	(0.0443)	(0.0330)	(0.0355)	(0.0308)	(0.0358)	(0.0320)	(0.0345)	(0.0334)
Male	0.0905	0.0711	0.0550	0.0721	0.0650	0.0538	0.0383	0.0807
	(0.0882)	(0.0684)	(0.0711)	(0.0612)	(0.0733)	(0.0656)	(0.0701)	(0.0655)
Enrollment Age	-0.244	-0.0173	-0.0494	0.00253	-0.0594	0.0291	-0.0635	0.0206
	(0.221)	(0.169)	(0.171)	(0.145)	(0.180)	(0.161)	(0.166)	(0.166)
HSGPA	-0.0350	-0.0961	-0.0569	-0.0308	-0.102	-0.0524	-0.0585	-0.0428
	(0.0776)	(0.0603)	(0.0612)	(0.0522)	(0.0642)	(0.0574)	(0.0603)	(0.0581)
Academic HS	-0.0519	-0.0479	-0.0236	-0.0227	-0.0267	-0.0286	-0.00804	-0.0231
	(0.0796)	(0.0625)	(0.0675)	(0.0587)	(0.0674)	(0.0606)	(0.0655)	(0.0629)
School Duration	-0.0320	0.00924	-0.00541	0.0543	-0.0101	0.0406	-0.00252	0.00762
	(0.137)	(0.105)	(0.111)	(0.0987)	(0.113)	(0.102)	(0.112)	(0.105)
No. of Students	392	727	727	1056	1056	1303	1303	1478
Polynomial Degree	1	1	2	2	3	3	4	4
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]

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

Note: OLS estimates using student background characteristics as outcome variables. The different columns use different estimation windows and varying polynomial orders of the underlying assignment variable. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

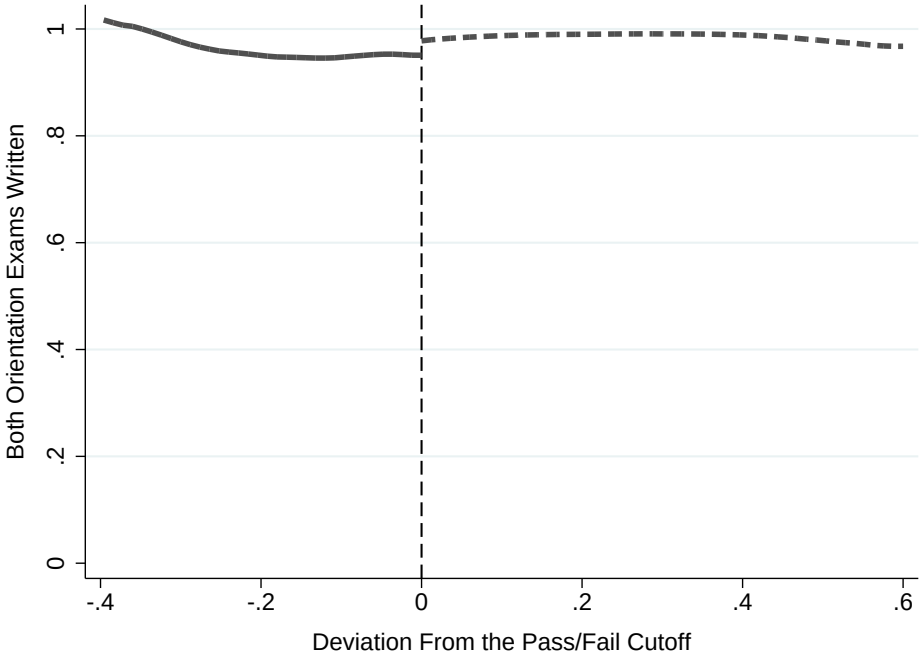
Figure 3.5: Average Grade After the First Semester



Note: Local polynomial functions are fitted to the left and the right of the cutoff based on Epanechnikov kernel. The optimal bandwidth is calculated according to Imbens and Kalyanaraman (2009).

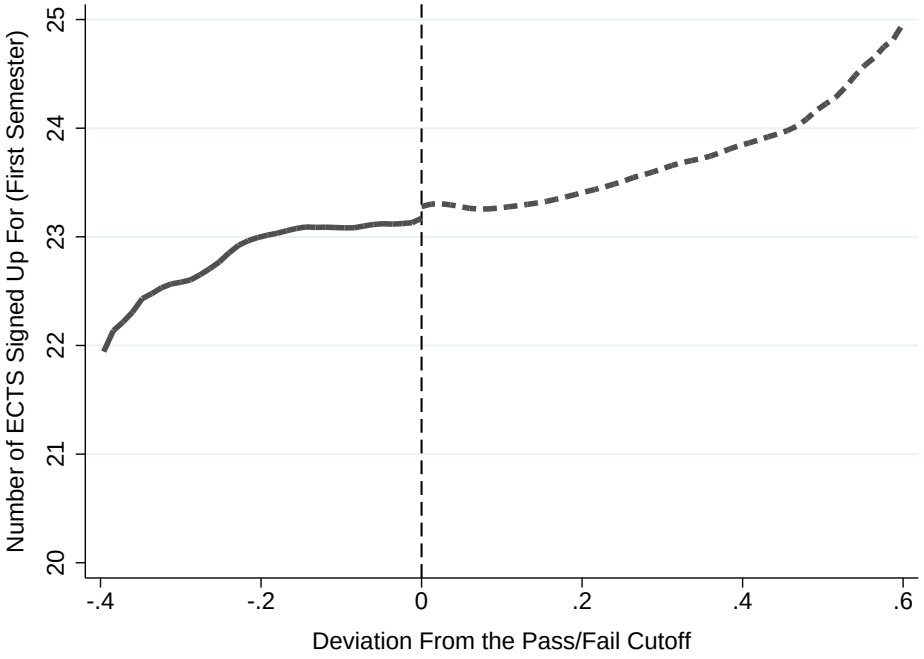
HSGPA is a very broad measure of ability and may not capture subject specific skills in economics, we further investigate whether students differ in their subject specific skills by inspecting graphically the average grade obtained in all exams written in the first semester at the first attempt besides Mathematics I. As depicted by Figure 3.5, there is virtually

Figure 3.6: Both Orientation Exams Written



Note: Local polynomial functions are fitted to the left and the right of the cutoff based on Epanechnikov kernel. The optimal bandwidth is calculated according to Imbens and Kalyanaraman (2009).

Figure 3.7: Number of ECTS Tried to Achieve in the First Semester



Note: Local polynomial functions are fitted to the left and the right of the cutoff based on Epanechnikov kernel. The optimal bandwidth is calculated according to Imbens and Kalyanaraman (2009).

no difference in the average grade obtained of students just above and just below the cutoff. Therefore, we conclude that there is no evidence for systematic sorting around the threshold, and our results can be interpreted as the effect of failing the first attempt of an orientation exam on marginal students' subsequent academic achievement.

A further concern regarding the validity of our results may be the argument made by Holmström and Milgrom (1991) in their theoretical paper on multitasking. While students are forced to write the orientation exams in the first semester, they are free to choose their workload besides these two courses. Therefore, it could be that students who narrowly failed an orientation exam are just worse in strategically planning the number of courses they attend. It could also be that students who narrowly passed an orientation exam stayed away from the first attempt with a medical certificate, thereby having more time for the exam preparation. We address both concerns by graphically investigating whether students just above and just below the cutoff are equally likely to have written both orientation exams, and whether they registered for the same number of exams in the first semester, at the first attempt. As shown in Figures 3.6 and 3.7, we find no significant difference in terms of the number of exams written of students at the cutoff. We conclude that strategic planning of exams written in the first semester does not explain our results.

3.4.4 Results

In the following, we present estimates of the effect of failing the first attempt of an orientation exam on marginal students' probability to drop-out and to graduate. Section 3.4.4 starts with the results of our baseline specifications on the probability to drop out. Afterwards, we examine the longer-term effect of failing an orientation exam on the probability to graduate. At the end of this section, we check the sensitivity of our baseline specification to changes in the sample restrictions and model specifications.

The Effect of Failing an Orientation Exam on Drop-Out

Students who narrowly failed an orientation exam are more likely to drop out of the program than students who narrowly passed the exam for several reasons. First, students who fail an orientation exam on their first attempt are likely to fail the exam on their second attempt as well, then being forced to drop out. Second, students who fail on their first attempt may decide to change the university (or the study program) before they lose the eligibility to study economics at any German public university. Third, students who fail may realize that passing the program would be very demanding for them and may decide to search for a job immediately. In the early phase of the program, we can therefore expect to find a significant difference in the drop-out rate between students who narrowly

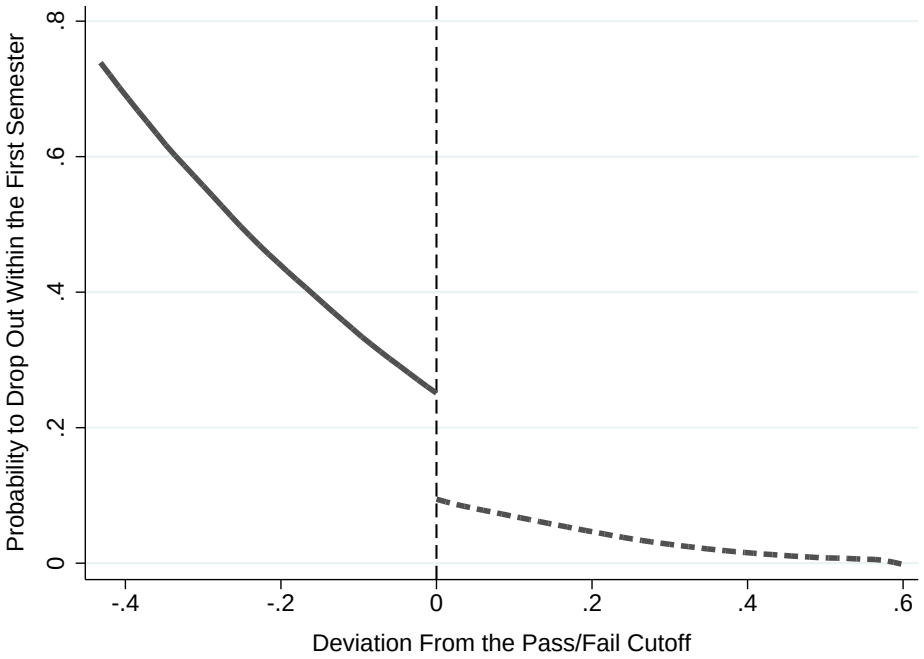
passed an orientation exam and students who narrowly failed it.

The development of this effect in the medium-term is not that easy to predict. However, as we eventually compare very similar students, we may observe a declining difference over time. The relative importance of failing or passing one specific exam at the first attempt, even if it is a high-stakes one, should decrease over time and with the number of exams written. Furthermore, students who narrowly fail an orientation exam may even gain a small competitive advantage over their fellow students, given they pass the retake exam, by having repeated the course material and potentially improved their learning strategies.

Figure 3.8 shows that the drop-out rate after the first semester of students who narrowly failed Mathematics I on their first attempt is indeed significantly higher than the drop-out rate of students who narrowly passed Mathematics I. While students who narrowly failed, dropped out with a probability of around 25 percent, only about 10 to 15 percent of the students who narrowly passed Mathematics I dropped out. The graph further shows that the relationship between the number of points achieved in an orientation exam and the drop-out probability after the first semester is negative on both sides of the cutoff. For very good students, drop-out probabilities are below 10 percent, while drop-out rates are around 50 percent at the left end of the distribution. Our RDD estimates reported in Table 3.9 confirm the pattern presented in Figure 3.8. The difference in the drop-out rates ranges from 16 to 19 percentage points across specifications and is always statistically significant at the one percent level. Considering that the drop-out probability of students who narrowly passed Mathematics I at the first attempt is only at around 10 to 15 percent, students who narrowly failed are more than twice as likely to drop out of the program after the first semester.

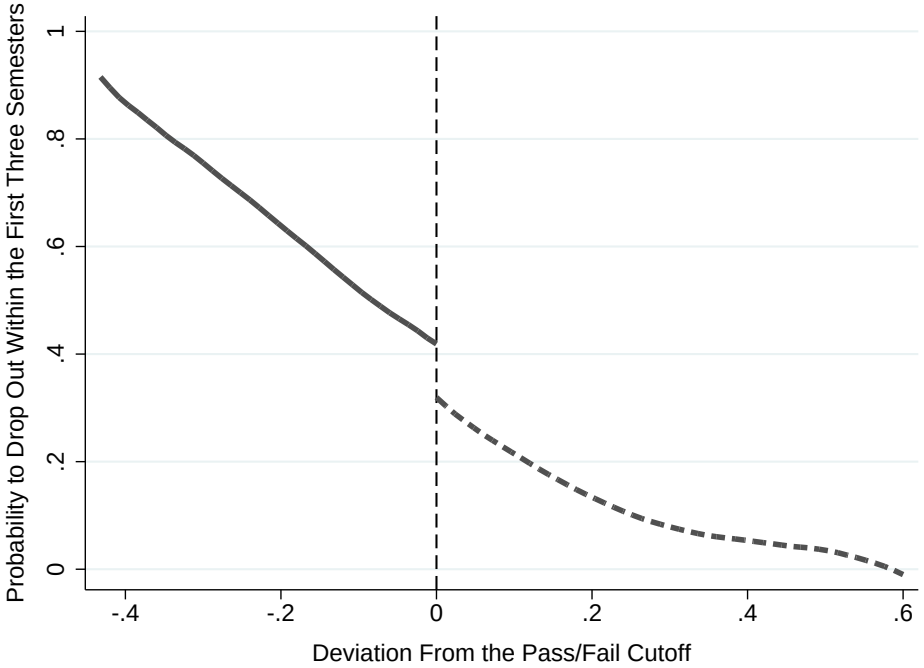
Investigating the effect on the drop-out rate at the end of the third semester, we find that the differential effect on the students just above and just below the cutoff persists, but becomes smaller. A visual illustration of the decline in the difference in the drop-out rate is provided by Figure 3.9. Although there is still a notable discontinuity at the cutoff, the jump is much smaller than the jump in Figure 3.8. The pattern found in Figures 3.8 and 3.9 is also confirmed by the estimates presented in the second line of Table 3.9. The difference in the drop-out rate between the students just above and just below the cutoff decreases to seven to twelve percentage points and is rarely statistically significant. Thus, we observe a decrease in the size of the effect of about 50 percent in comparison to end of the first semester, which is also in line with our prediction. This implies that – after the first semester – students who narrowly pass an orientation exam are more likely to drop out of the program than students who narrowly fail an orientation exam, given they pass the exam on their second attempt. This result stresses the positive effect of high-stakes exams for 'surviving' students.

Figure 3.8: The Effect on Drop-out After the First Semester



Note: Local polynomial functions are fitted to the left and the right of the cutoff based on Epanechnikov kernel. The optimal bandwidth is calculated according to Imbens and Kalyanaraman (2009).

Figure 3.9: The Effect on Drop-out After the Third Semster



Note: Local polynomial functions are fitted to the left and the right of the cutoff based on Epanechnikov kernel. The optimal bandwidth is calculated according to Imbens and Kalyanaraman (2009).

Table 3.9: RDD Estimates: The Effect of Failing an Orientation Exam on Drop-Out

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Drop-out I	0.188***	0.193***	0.179***	0.163***	0.189***	0.171***	0.182***	0.164***	0.163***
	(0.0648)	(0.0492)	(0.0576)	(0.0504)	(0.0537)	(0.0477)	(0.0566)	(0.0532)	(0.051)
Drop-out III	0.0963	0.0988	0.102	0.104*	0.0951	0.0756	0.0916	0.0886	0.108
	(0.0838)	(0.0632)	(0.0689)	(0.0589)	(0.0683)	(0.0607)	(0.0664)	(0.0627)	(0.067)
#Students	392	727	727	1056	1056	1303	1303	1478	1478
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

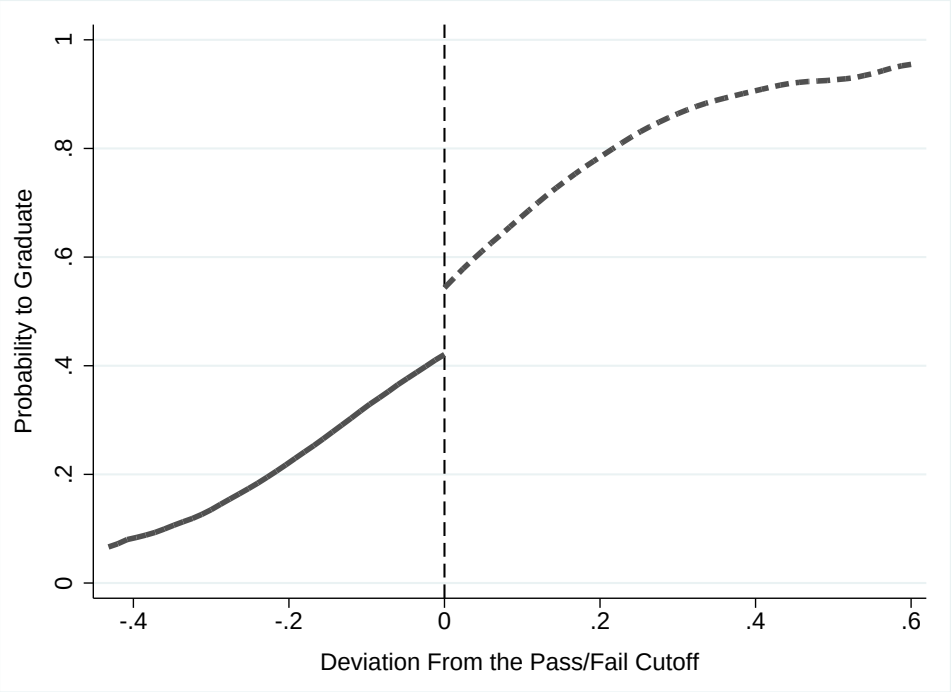
Note: Linear probability models and local linear regressions. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. All parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

The Effect of Failing an Orientation Exam on Graduation

Repeating the course material and spending some additional time on the preparation of the exam may have a positive effect on academic achievement in later semesters. Students may fail an exam, e.g., because they miss the required skills, or because they may have problems to adopt to the high level of self-motivated learning. The second attempt gives students who failed an orientation exam the chance to improve their skills, and to put more effort into the exam preparation. Since students just above and below the cutoff have very similar or almost identical characteristics, the additional effort should have a positive effect on their subsequent performance, which may also affect their probability to graduate.

A visual illustration of the relationship between our assignment variable and the probability to graduate is provided by Figure 3.10. The discontinuity found at the cutoff is highly comparable to the discontinuity found in Figure 3.9. Our RDD estimates, reported in Table 3.10, confirm the graphical inspection. In all specifications, we find a negative, though not constantly significant effect of being exposed to the retake exam on the probability to graduate. Thus, failing the first attempt decreases students' probability to graduate by about nine to thirteen percentage points, which is quantitatively comparable to the results we obtained for drop-out until the end of the third semester (see Table 3.9). This result suggests that – after the first half of the program – the treatment has no significant effect anymore on marginal students' drop-out probability. Furthermore, it indicates that the positive effect of failing the first attempt in terms of learning advantages is not strong enough to compensate completely for the strong (mechanical) effect until the end of the first semester. Therefore, highly comparable students are eventually less likely to graduate if they by chance fail the first attempt of an orientation exam and are exposed to the high-stakes exam.

Figure 3.10: The Effect on Graduation



Note: Local polynomial functions are fitted to the left and the right of the cutoff based on Epanechnikov kernel. The optimal bandwidth is calculated according to Imbens and Kalyanaraman (2009).

Table 3.10: RDD Estimates: The Effect of Failing an Orientation Exam on Graduation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Graduation	-0.131	-0.0960	-0.115*	-0.126**	-0.104	-0.0929	-0.108	-0.103	-0.134*
	(0.0866)	(0.0657)	(0.0690)	(0.0588)	(0.0706)	(0.0629)	(0.0668)	(0.0633)	(0.069)
#Students	392	727	727	1056	1056	1303	1303	1478	1478
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. *Graduation* is a dummy variable equal to one if a student graduated. All parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

Robustness Checks

In the following, we check the sensitivity of our baseline results to several modifications and different sample definitions, demonstrating that our findings hold. In particular, we check whether the results differ if we consider a different orientation exam, perform placebo tests, vary the underlying polynomial function, or the sample considered.

To verify that the high-stakes testing mechanism explains our results, we first use the deviation from the pass/fail cutoff relative to the maximum number of points achieved in the second orientation exam of the first semester, i.e., in Introduction to Economics, as our assignment variable. The results presented in Tables 3.A.1 and 3.A.2 are very similar

to our baseline results. Students who narrowly failed the Introduction to Economics exam at the first attempt are 10 to 17 percentage points more likely to drop out until the end of the first semester than students who narrowly passed. The effect on drop-out until the end of the third semester is in comparison to our baseline result slightly stronger and more significant, which may be explained by varying grading standards. An exam with a lower grading standard may, on the one hand, have a smaller effect until the end of the first semester, as students who fail have a higher chance to survive the second attempt, but may, on the other hand, be less beneficial in helping marginal students to catch up with their better performing classmates and to adopt to courses with higher standards. However, until graduation, the effect becomes again highly comparable to our baseline result: Students who failed the Introduction to Economics exam on their first attempt are six to twelve percentage points less likely to graduate, though this effect is in only one specification statistically significant.

If the effects on marginal students' academic achievement are causal effects of high-stakes testing, we should not find any significant effects if we use the deviation from the pass/fail cutoff in a non-orientation exam of the first semester. Table 3.A.3 shows results of our baseline specifications using the deviation from the pass/fail cutoff relative to the maximum number of points in the Introduction to Accounting exam as our assignment variable. This exam is one of the three non-orientation exams in the first semester. The estimates show no significant effect with respect to any of the outcome variables, suggesting that our baseline estimates reflect the causal effect of the high-stakes testing mechanism.

A further concern may be that the assignment variable in our baseline specifications is fundamentally discontinuous, and the estimated effects are explained by other factors than the treatment. To test whether the treatment effect is zero when it should be, we conduct two placebo test using cutoffs at 10 and 20 points above the actual pass/fail cutoff in Mathematics I. Jumps at placebo cutoffs would not invalidate our baseline results, but require an explanation. However, as shown in Tables 3.A.4 and 3.A.5, we find no evidence for a discontinuity at the placebo cutoffs.

In our baseline specifications, we assume the parameters of the polynomial function to be identical on both sides of the cutoff. To test the sensitivity of our baseline results, we now allow the parameters to differ to the right and to the left of the cutoff. The results are reported in Table 3.A.6. We still find that failing the first attempt significantly increases the probability to drop out until the end of the first semester, while the effect on graduation is also negative but slightly less statistically significant.

Finally, we check whether our result regarding the graduation rate changes if we exclude the last cohort. A concern may be that students who fail the first attempt lose time, as they have to repeat the orientation exam, and therefore graduate at a later point in time

compared to the students in the control group. As we observe the last cohort for 'only' eight semesters, we now investigate whether our results hold if we follow students for at least 10 semesters by excluding the last cohort from our sample. The results are reported in Table 3.A.7. We still find a negative effect on the probability to graduate of eight to fifteen percentage points.

3.5 Discussion and Conclusion

This article investigates the effect of a high-stakes testing policy, introduced in a large German federal state in the early 2000's, on undergraduates' academic achievement. Aiming to reduce late drop-out, students affected by the policy have to pass so-called orientation exam(s) until the end of their third semester to be allowed to continue their studies. Students who fail an orientation exam twice lose the eligibility to study their subject at every German university.

Using administrative panel data on all university students in Germany from 1997 to 2003, we assess, in the first part of this paper, the average causal effect of the introduction of orientation exams on drop-out within the first two years. Exploiting variation between states over time allows us to estimate DiD models. Our results show that students affected by the reform where on average about 10 percent more likely to drop out of their program until the end of the second year. To further investigate the mechanism and the effectiveness of these high-stakes exams, we use in the second part of this paper administrative, student-level panel data of the Economics program at the University of Konstanz. We account for the endogeneity of the treatment status in a regression discontinuity framework, thus obtaining local estimates of the effect of failing an orientation exam in the first semester on dropout and graduation probabilities. We find that failing an orientation exam increases the probability to drop out of the program after the first semester by 16 to 19 percentage points. Over time, the difference in the drop-out rate between treatment and control students becomes smaller but remains significant until graduation.

Our results suggest that high-stakes testing at an early stage of a study program can help to reduce the problem of late drop-outs. However, our RDD estimates also show the difficulty in finding the optimal grading standard in an high-stakes exam. Although having approximately the same ability, students who just fail to achieve the required threshold in an orientation exam have a lower probability to graduate than students who just pass the threshold. Policies similar to the one implemented in North Rhine-Westphalia, where universities receive a lump sum for every graduate, would additionally compromise the effectiveness of orientation exams in reducing late drop-out as they create an incentive for lower grading standards, especially in higher semesters of a program.

References

Alexander, F. K. (2000). The changing face of accountability: Monitoring and assessing institutional performance in higher education. *The Journal of Higher Education*, 71(4):411–431.

Alexander, K., Entwisle, D., and Kabbani, N. (2003). Grade retention, social promotion, and 'third way' alternatives. *Early Childhood Programm for a New Century*, pages 197–238.

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

Angrist, J. and Pischke, J. (2009). *Mostly Harmless Econometrics: An Empirical Companion*. Princeton University Press.

Arnold, I. J. (2015). The effectiveness of academic dismissal policies in Dutch university education: An empirical investigation. *Studies in Higher Education*, 40(6):1068–1084.

Betts, J. R. and Costrell, R. M. (2001). Incentives and equity under standards-based reform. *Brookings Papers on Education Policy*, 4(4):9–74.

Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *Journal of Economic Education*, 29(2):171–182.

Bölke, L. and Haug, V. (2009). *Das Hochschulrecht in Baden-Württemberg: Eine systematische Darstellung*. C. F. Müller Wissenschaft. Müller.

Carnoy, M. and Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4):305–331.

Clark, D. and Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, 122(2):282 – 318.

Conner, T. W. and Rabovsky, T. M. (2011). Accountability, affordability, access: A review of the recent trends in higher education policy research. *Policy Studies Journal*, 39(s1):93–112.

Geiser, S. and Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. *Research & Occasional Paper Series: CSHE.6.07.*

Gerard, F., Rokkanen, M., and Rothe, C. (2015). Identification and inference in regression discontinuity designs with a manipulated running variable. IZA Discussion Papers 9604, Bonn.

Hanushek, E. A. and Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2):297–327.

Hillman, N. W., Tandberg, D. A., and Fryar, A. H. (2015). Evaluating the impacts of new performance funding in higher education. *Educational Evaluation and Policy Analysis*, 37(4):501–519.

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

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, 23(2):99–121.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89(5-6):761–796.

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244.

Martorell, F. (2004). Do high school graduation exams matter? A regression discontinuity approach. Department of Economics, UC Berkeley.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Neill, M. and Gayler, K. (2001). Do high-stakes graduation tests improve learning outcomes? Using state-level NAEP data to evaluate the effects of mandatory graduation tests. In Orfield, G. and Kornhaber, M. L., editors, *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*, pages 107–126. New York: Century Foundation Press.

OECD (2010). *Education at a Glance 2010*. OECD Indicators, Paris. OECD Publishing.

Rothstein, J. (2004). College performance predictions and the SAT. *Journal of Econometrics*, 121(1-2):297–317.

Tafreschi, D. and Thiemann, P. (2016). Doing it twice, getting it right? The effects of grade retention and course repetition in higher education. *Economics of Education Review*, 55:198 – 219.

Trapmann, S., Hell, B., Weigand, S., and Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs : eine Metaanalyse. *Zeitschrift für pädagogische Psychologie*, 21(1):11–27.

Urquiola, M. and Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, 99(1):179–215.

Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., and Wollscheid, S. (2015). Dropout and completion in higher education in Europe: Main report. Technical report, Luxembourg: Publications Office of the European Union.

3.6 Appendix

Table 3.A.1: Robustness Check: The Effect of Failing Introduction to Economics on Drop-Out

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Drop-out I	0.123**	0.166***	0.139**	0.135***	0.147***	0.157***	0.103*	0.115**	0.1302**
	(0.0647)	(0.0467)	(0.0550)	(0.0479)	(0.0521)	(0.0457)	(0.0542)	(0.0524)	(0.0598)
Drop-out III	0.161**	0.126**	0.132**	0.0994*	0.161***	0.125**	0.101	0.106*	0.1261*
	(0.0434)	(0.0564)	(0.0625)	(0.0531)	(0.0619)	(0.0542)	(0.0616)	(0.0602)	(0.0672)
#Students	538	1005	1005	1358	1358	1572	1572	1695	1695
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. Parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

Table 3.A.2: Robustness Check: The Effect of Failing Introduction to Economics on Graduation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Graduation	-0.0896	-0.0837	-0.0835	-0.0625	-0.122*	-0.0841	-0.0550	-0.0620	-0.1011
	(0.0683)	(0.0594)	(0.0627)	(0.0524)	(0.0648)	(0.0564)	(0.0621)	(0.0613)	(0.0692)
#Students	538	1005	1005	1358	1358	1572	1572	1695	1695
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. *Graduation* is a dummy variable equal to one if a student graduated. Parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

Table 3.A.3: Robustness Check: The Effect of Failing a Non-Orientation Exam

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Drop-out I	0.0113	0.0386	0.0281	0.0222	0.0155	0.0104	0.00464	-0.00450	0.0018
	(0.0457)	(0.0347)	(0.0380)	(0.0331)	(0.0370)	(0.0334)	(0.0389)	(0.0348)	(0.0387)
Drop-out III	0.103	0.0886*	0.0865	0.0603	0.0812	0.0610	0.0806	0.0308	0.0815
	(0.0719)	(0.0509)	(0.0553)	(0.0470)	(0.0556)	(0.0491)	(0.0555)	(0.0501)	(0.0625)
Graduation	-0.0710	-0.0646	-0.0776	-0.0673	-0.0724	-0.0533	-0.0741	-0.0271	-0.0412
	(0.0815)	(0.0574)	(0.0610)	(0.0514)	(0.0627)	(0.0554)	(0.0609)	(0.0551)	(0 .0701)
#Students	485	973	973	1326	1326	1511	1511	1559	1559
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. The assignment variable is the deviation from the pass/fail cutoff, relative to the maximum number of points, achieved at the first attempt of the Introduction to Accounting exam. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. *Graduation* is a dummy variable equal to one if a student graduated. Parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable. Cohort fixed effects are included.

Table 3.A.4: Robustness Check: Placebo Cutoff I

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Drop-out I	-0.00503	0.0302	0.0302	-0.0136	0.0374	0.0134	0.0169	0.00796	0.011
	(0.0315)	(0.0237)	(0.0239)	(0.0214)	(0.0278)	(0.0280)	(0.0241)	(0.0268)	(0.0236)
Drop-out III	0.0371	0.0644	0.0631	0.00297	0.0632	0.0209	0.0366	0.0283	0.032
	(0.0530)	(0.0392)	(0.0394)	(0.0332)	(0.0435)	(0.0413)	(0.0382)	(0.0393)	(0.036)
Graduation	-0.0172	-0.0226	-0.0202	-0.00506	-0.0307	0.00102	-0.0172	-0.0120	-0.0354
	(0.0695)	(0.0501)	(0.0503)	(0.0418)	(0.0555)	(0.0505)	(0.0492)	(0.0489)	(0.059)
#Students	457	874	874	1207	1207	1333	1333	1478	1478
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. *Graduation* is a dummy variable equal to one if a student graduated. Parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

Table 3.A.5: Robustness Check: Placebo Cutoff II

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Drop-out I	-0.0383	-0.0197	-0.0220	-0.0159	-0.0179	-0.0221	-0.0104	-0.0263	-0.0304
	(0.0321)	(0.0249)	(0.0279)	(0.0279)	(0.0276)	(0.0268)	(0.0293)	(0.0294)	(0.031)
Drop-out III	0.0220	-0.00013	0.00229	0.0263	0.0205	0.00353	0.00232	0.00041	-0.0026
	(0.0634)	(0.0432)	(0.0468)	(0.0485)	(0.0473)	(0.0447)	(0.0529)	(0.0462)	(0.0498)
Graduation	-0.00557	0.0245	0.0486	0.0371	0.0426	0.0616	0.0628	0.0509	0.052
	(0.0894)	(0.0597)	(0.0684)	(0.0679)	(0.0669)	(0.0648)	(0.0723)	(0.0643)	(0.070)
#Students	213	424	424	678	678	907	907	1478	1478
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. *Graduation* is a dummy variable equal to one if a student graduated. Parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

Table 3.A.6: Robustness Check: Baseline Estimates Using Asymmetric Polynomials

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Drop-out I	0.199***	0.175***	0.170**	0.183**	0.149	0.166**	0.147	0.132	0.152
	(0.0803)	(0.0595)	(0.0856)	(0.0723)	(0.0943)	(0.0835)	(0.104)	(0.0992)	(0.0969)
Drop-out III	0.162*	0.107	0.146	0.102	0.158	0.125	0.156	0.154	0.156
	(0.0948)	(0.0700)	(0.100)	(0.0842)	(0.110)	(0.0963)	(0.120)	(0.114)	(0.112)
Graduation	-0.158*	-0.120*	-0.169*	-0.115	-0.160	-0.143	-0.165	-0.167	-0.167
	(0.0955)	(0.0699)	(0.0992)	(0.0839)	(0.109)	(0.0961)	(0.119)	(0.114)	(0.111)
#Students	392	727	727	1056	1056	1303	1303	1478	1478
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. *Graduation* is a dummy variable equal to one if a student graduated. Parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

Table 3.A.7: Robustness Check: Baseline Estimates, Last Cohort Excluded

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Drop-out I	0.250***	0.231***	0.228***	0.197***	0.231***	0.198***	0.225***	0.194***	0.221***
	(0.0808)	(0.0594)	(0.0699)	(0.0598)	(0.0651)	(0.0570)	(0.0677)	(0.0637)	(0.0677)
Drop-out III	0.132	0.117	0.135*	0.119*	0.119	0.0717	0.118	0.0971	0.125*
	(0.101)	(0.0751)	(0.0816)	(0.0686)	(0.0804)	(0.0712)	(0.0774)	(0.0731)	(0.0694)
Graduation	-0.174*	-0.104	-0.149*	-0.131*	-0.123	-0.0817	-0.126*	-0.109	-0.157
	(0.100)	(0.0757)	(0.0792)	(0.0668)	(0.0805)	(0.0717)	(0.0757)	(0.0717)	(0.0760)
#Students	304	563	563	825	825	1036	1036	1190	1190
Polynomial	1	1	2	2	3	3	4	4	NP
Window	[-0.1,0.1]	[-0.2,0.2]	[-0.2,0.2]	[-0.3,0.3]	[-0.3,0.3]	[-0.4,0.4]	[-0.4,0.4]	[.]	[.]

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

Note: Linear probability models and local linear regressions. *Drop-out I* and *Drop-out III* are dummy variables equal to one if a student dropped out after the first, or within the first three semesters, respectively. *Graduation* is a dummy variable equal to one if a student graduated. All parametric regressions control for gender, age, nationality, HSGPA, high school type attended, and cohort fixed effects. Standard Errors reported in parentheses are clustered at the level of the assignment variable.

Complete Bibliography

Alexander, F. K. (2000). The changing face of accountability: Monitoring and assessing institutional performance in higher education. *The Journal of Higher Education*, 71(4):411–431.

Alexander, K., Entwisle, D., and Kabbani, N. (2003). Grade retention, social promotion, and 'third way' alternatives. *Early Childhood Programm for a New Century*, pages 197–238.

Almas, I., Cappelen, A. W., Salvanes, K. G., Sorensen, E. O., and Tungodden, B. (2016). Willingness to compete: Family matters. *Management Science*, 62(8):2149–2162.

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Chapter 1 - Personality psychology and economics. In Eric A. Hanushek, S. M. and Wößmann, L., editors, *Handbook of The Economics of Education*, volume 4 of *Handbook of the Economics of Education*, pages 1 – 181. Elsevier.

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

Amabile, T. (1996). *Creativity In Context: Update To The Social Psychology Of Creativity*. Westview Press.

Anderson, D. M. and Walker, M. (2015). Does shortening the school week impact student performance? Evidence from the four-day school week. *Education Finance and Policy*, 10(3):314–349.

Andrietti, V. (2015). The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment. Universidad Carlos III de Madrid Working Paper Economic Series 15-06.

Angrist, J. and Pischke, J. (2009). *Mostly Harmless Econometrics: An Empirical Companion*. Princeton University Press.

Arnold, I. J. (2015). The effectiveness of academic dismissal policies in Dutch university education: An empirical investigation. *Studies in Higher Education*, 40(6):1068–1084.

Autorengruppe Bildungsberichterstattung, editor (2008). *Bildung in Deutschland 2008. Ein indikatorengesttzter Bericht mit einer Analyse zu bergngen im Anschluss an den Sekundarbereich I.* Bertelsmann, Bielefeld.

Autorengruppe Bildungsberichterstattung, editor (2010). *Bildung in Deutschland 2010. Ein indikatorengesttzter Bericht mit einer Analyse zu Perspektiven des Bildungswesens im demografischen Wandel.* Bertelsmann, Bielefeld.

Battistin, E. and Meroni, E. C. (2016). Should we increase instruction time in low achieving schools? Evidence from Southern Italy. *Economics of Education Review*, 55:39 – 56.

Baumert, J., Maaz, K., Gresch, C., McElvany, N., Anders, Y., Jonkmann, K., Neumann, M., and Watermann, R. (2010). *Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten. Zusammenfassung der zentralen Befunde.* Bundesministerium für Bildung und Forschung, Referat Bildungsforschung, Bonn.

Becker, W. and Rosen, S. (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review*, 11(2):107–118.

Bellei, C. (2009). Does lengthening the school day increase students academic achievement? Results from a natural experiment in Chile. *Economics of Education Review*, 28(5):629–640.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Betts, J. R. (1998). The impact of educational standards on the level and distribution of earnings. *The American Economic Review*, 88(1):266–275.

Betts, J. R. and Costrell, R. M. (2001). Incentives and equity under standards-based reform. *Brookings Papers on Education Policy*, 4(4):9–74.

Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *Journal of Economic Education*, 29(2):171–182.

Bölke, L. and Haug, V. (2009). *Das Hochschulrecht in Baden-Württemberg: Eine systematische Darstellung.* C. F. Müller Wissenschaft. Müller.

Booth, A. and Nolen, P. (2012). Choosing to compete: How different are girls and boys? *Journal of Economic Behavior & Organization*, 81(2):542 – 555.

Büttner, B. and Thomsen, S. L. (2013). Are we spending too many years in school? Causal evidence of the impact of shortening secondary school duration. *German Economic Review*, 16(1):65–86.

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education*, 78(1):27–49.

Card, D. (1999). The Causal Effect of Education on Earnings. In Ashenfelter, O. C. and Card, D., editors, *Handbook of Labor Economics*, volume 3, pages 1801 – 1863. Amsterdam: Elsevier.

Carneiro, P., Crawford, C., and Goodman, A. (2007). The impact of early cognitive and non-cognitive skills on later outcomes. CEE Discussion Papers, Centre for the Economics of Education, LSE.

Carnoy, M. and Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4):305–331.

Chevalier, J. and Ellison, G. (1999). Career concerns of mutual fund managers. *The Quarterly Journal of Economics*, 114(2):389–432.

Clark, D. and Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, 122(2):282 – 318.

Conner, T. W. and Rabovsky, T. M. (2011). Accountability, affordability, access: A review of the recent trends in higher education policy research. *Policy Studies Journal*, 39(s1):93–112.

Dahmann, S. and Anger, S. (2014). The impact of education on personality: Evidence from a German high school reform. SOEPPapers on Multidisciplinary Panel Data Research 658, DIW Berlin.

Dahmann, S. C. (2017). How does education improve cognitive skills? Instructional time versus timing of instruction. *Labour Economics*.

Dechenaux, E., Kovenock, D., and Sheremeta, R. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

Eisenkopf, G. (2009). Student selection and incentives. *Zeitschrift für Betriebswirtschaft*, 79(5):563–577.

Eisenkopf, G. and Teyssier, S. (2013). Envy and loss aversion in tournaments. *Journal of Economic Psychology*, 34(C):240–255.

Eren, O. and Millimet, D. L. (2007). Time to learn? The organizational structure of schools and student achievement. *Empirical Economics*, 32(2-3):301–332.

Fama, E. (1980). Agency problems and the theory of the firm. *Journal of Political Economy*, 88(2):288–307.

Federal Statistical Office, editor (2010). *Wirtschaft und Statistik*. Federal Statistical Office, Wiesbaden.

Fischer, F., Schult, J., and Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*, 28(2):529–543.

Frey, B. and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5):589–611.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review*, 87(4):746–755.

Geiser, S. and Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. *Research & Occasional Paper Series: CSHE.6.07.*

Gerard, F., Rokkanen, M., and Rothe, C. (2015). Identification and inference in regression discontinuity designs with a manipulated running variable. IZA Discussion Papers 9604, Bonn.

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

Guay, F., Ratelle, C. F., and Chanal, J. (2008). Optimal learning in optimal contexts: The role of self-determination in education. *Canadian Psychology*, 49(3):233.

Hansen, B. (2011). School year length and student performance: Quasi-experimental evidence. Technical report.

Hanushek, E., Rivkin, S., and Taylor, L. L. (1996). Aggregation and the estimated effects of school resources. *The Review of Economics and Statistics*, 78(4):611–27.

Hanushek, E. A. and Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2):297–327.

Hanushek, E. A. and Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences- in-differences evidence across countries. *Economic Journal*, 116(510):C63–C76.

Hanushek, E. A. and Wößmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3):607–68.

Harris, M. and Holmström, B. (1982). A theory of wage dynamics. *The Review of Economic Studies*, 49(3):315–333.

Hillman, N. W., Tandberg, D. A., and Fryar, A. H. (2015). Evaluating the impacts of new performance funding in higher education. *Educational Evaluation and Policy Analysis*, 37(4):501–519.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182.

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

Homuth, C. (2012). Der Einfluss des achtjährigen Gymnasiums auf den Kompetenzerwerb. Technical report.

Huebener, M. and Marcus, J. (2017). Compressing instruction time into fewer years of schooling and the impact on student performance. *Economics of Education Review*, 58:1 – 14.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

ISB Staatsinstitut für Schulqualität und Bildungsforschung, editor (2009). *Bildungsbericht Bayern 2009*. Kastner AG, Wolnzach.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, 23(2):99–121.

Jacob, B. A. (2005a). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89(5-6):761–796.

Jacob, B. A. (2005b). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89(5-6):761–796.

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244.

Jürges, H., Schneider, K., Senkbeil, M., and Carstensen, C. H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31(1):56–65.

Klemm, K. (2008). Bildungszeit: Vom Umgang mit einem knappen Gut. In *Schulzeiten, Lernzeiten, Lebenszeiten*, pages 21–30. Zeiher, H./ Schroeder, S., Weinheim.

Koerselman, K. (2013). Incentives from curriculum tracking. *Economics of Education Review*, 32:140–150.

Krashinsky, H. (2014). How would one extra year of high school affect academic performance in university? Evidence from an educational policy change. *Canadian Journal of Economics*, 47(1):70–97.

Kühn, S. M., van Ackeren, I., Bellenberg, G., Reintjes, C., and im Brahm, J.-P. D. G. (2013). Wie viele Schuljahre bis zum Abitur? *Zeitschrift für Erziehungswissenschaft*, 16(1):115–136.

Lavy, V. (2012). Expanding school resources and increasing time on task: Effects of a policy experiment in Israel on student academic achievement and behaviour. NBER Working Paper 18369.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, 125(588):F397–F424.

Lazear, E. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–64.

Lee, J.-W. and Barro, R. J. (2001). Schooling quality in a cross–section of countries. *Economica*, 68(272):465–488.

Lehn, B. v. (2010). *Generation G8. Wie die Turbo-Schule Schüler und Familien ruiniert.* Weinheim: Beltz.

Lochner, L. (2011). Nonproduction benefits of education: Crime, health, and good citizenship. *Handbook of the Economics of Education*, 4:183.

Malamud, O. and Pop-Eleches, C. (2011). School tracking and access to higher education among disadvantaged groups. *Journal of Public Economics*, 95(11-12):1538–1549. Special Issue: International Seminar for Public Economics on Normative Tax Theory.

Mandel, P. and Süssmuth, B. (2011). Total instructional time exposure and student achievement: An extreme bounds analysis based on German state-level variation. CESifo Working Paper Series 3580.

Marcotte, D. E. (2007). Schooling and test scores: A mother-natural experiment. *Economics of Education Review*, 26(5):629–640.

Marcotte, D. E. and Hemelt, S. W. (2008). Unscheduled school closings and student performance. *Education Finance and Policy*, 3(3):316–338.

Martorell, F. (2004). Do high school graduation exams matter? A regression discontinuity approach. Department of Economics, UC Berkeley.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Meghir, C. and Palme, M. (2005). Educational reform, ability, and family background. *The American Economic Review*, 95(1):414–424.

Meyer, T. and Thomsen, S. L. (2013). Are 12 years of schooling sufficient preparation for tertiary education? Evidence from the reform of secondary school duration in Germany. NIW Discussion Paper 8.

Meyer, T. and Thomsen, S. L. (2016). How important is secondary school duration for postsecondary education decisions? Evidence from a natural experiment. *Journal of Human Capital*, 10(1):67–108.

Meyer, T., Thomsen, S. L., and Schneider, H. (2015). New Evidence on the Effects of the Shortened School Duration in the German States: An Evaluation of Post-Secondary Education Decisions. IZA Discussion Papers 9507, Institute for the Study of Labor (IZA).

Morin, L.-P. (2013). Estimating the benefit of high school for universitybound students: Evidence of subjectspecific human capital accumulation. *Canadian Journal of Economics*, 46(2):441–468.

Mudiappa, M. and Artelt, C. (2014). *BiKS - Ergebnisse aus den Längsschnittstudien. Praxisrelevante Befunde aus dem Primar- und Sekundarschulbereich.* University of Bamberg Press, Bamberg.

Neill, M. and Gayler, K. (2001). Do high-stakes graduation tests improve learning outcomes? Using state-level NAEP data to evaluate the effects of mandatory graduation

tests. In Orfield, G. and Kornhaber, M. L., editors, *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*, pages 107–126. New York: Century Foundation Press.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

OECD (2010). *Education at a Glance 2010*. OECD Indicators, Paris. OECD Publishing.

OECD (2014). *Education at a Glance 2014: OECD Indicators*. OECD Publishing.

Patall, E. A., Cooper, H., and Allen, A. B. (2010). Extending the school day or school year a systematic review of research (1985–2009). *Review of Educational Research*, 80(3):401–436.

Pischke, J.-S. (2007). The impact of length of the school year on student performance and earnings: Evidence from the German short school years. *The Economic Journal*, 117(523):1216–1242.

Pischke, J.-S. and Manning, A. (2006). Comprehensive versus selective schooling in England in Wales: What do we know? Working Paper 12176, National Bureau of Economic Research.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63.

Quis, J. S. and Reif, S. (2017). Health effects of instruction intensity. Evidence from a natural experiment in german high-schools. BERG Working Paper Series 123, Bamberg University.

Rosen, S. (1982). Authority, control, and the distribution of earnings. *Bell Journal of Economics*, 13(2):311–323.

Rosenbaum, J. (1976). *Making inequality: The hidden curriculum of high school tracking*. Wiley-interscience publication. Wiley, New York.

Rosenbaum, J. (1984). *Career mobility in a corporate hierarchy*. Academic Press, Orlando, FL.

Rosenbaum, J. E. (1979). Tournament mobility: Career patterns in a corporation. *Administrative Science Quarterly*, 24(2):220–241.

Rothstein, J. (2004). College performance predictions and the SAT. *Journal of Econometrics*, 121(1-2):297–317.

Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, pages 68–78.

Schnepf, S. V. (2003). Inequalities in secondary school attendance in germany.

Sirin, S. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3):417–453.

Smerdon, B. (1999). Engagement and achievement: Differences between african-american and white high school students. *Research in Sociology of Education and Socialization*, 12:103–34.

Soskice, D. (1994). Reconciling markets and institutions: The german apprenticeship system. In *Training and the Private Sector: International Comparisons*, pages 25–60. National Bureau of Economic Research, Inc.

Statistisches Bundesamt (2014). Bildung und Kultur: Allgemeinbildende Schulen. Fachserie 11 Reihe 1, Statistisches Bundesamt, Wiesbaden.

Tafreschi, D. and Thiemann, P. (2016). Doing it twice, getting it right? The effects of grade retention and course repetition in higher education. *Economics of Education Review*, 55:198 – 219.

Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A., and Updegraff, J. A. (2000). Biobehavioral Responses to Stress in Females: Tend-and-Befriend, Not Fight-or-Flight. *Psychological Review*, 107(3):411–429.

Thiel, H., Thomsen, S. L., and Büttner, B. (2014). Variation of learning intensity in late adolescence and the effect on personality traits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(4):861–892.

Trapmann, S., Hell, B., Weigand, S., and Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs : eine Metaanalyse. *Zeitschrift für pädagogische Psychologie*, 21(1):11–27.

Urquiola, M. and Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, 99(1):179–215.

Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., and Wollscheid, S. (2015). Dropout and completion in higher education in Europe: Main report. Technical report, Luxembourg: Publications Office of the European Union.

Winkelmann, R. (1996). Employment prospects and skill acquisition of apprenticeship-trained workers in germany. *Industrial and Labor Relations Review*, 49(4):658–672.

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2):117–170.

Abgrenzung

Ich versichere hiermit, dass ich das erste Kapitel "Tournaments at School: The Incentive Effect of Tracking on Student Effort and Skill Development" selbstständig und ohne Hilfe Dritter und ohne Benutzung anderer als der angegeben Hilfemittel erstellt habe.

Das zweite Kapitel "The Effect of a Compressed High School Curriculum on University Performance" ist in Zusammenarbeit mit Verena Lauber entstanden. Meine individuelle Leistung bei der Erstellung dieser Arbeit beträgt mindestens 50 Prozent.

Das dritte Kapitel "Accountability in Higher Education: The Impact of High-Stakes Testing on Academic Achievement" ist in Zusammenarbeit mit Enzo Brox entstanden. Meine individuelle Leistung bei der Erstellung dieser Arbeit beträgt mindestens 50 Prozent.